SEQUENCE MATCHING, SIMPLE SEARCHING

PGA Course in Bioinformatics
Tools for Comparative Analysis

May 12, 2003

Outline

- Sequence alignment algorithms
  - Rigorous Optimality:Needleman-Wunsch and Smith-Waterman
  - Rapid, heuristic algorithms
    - BLAST
    - FASTA
    - and their relatives
- Databases and Search Tools

What do you want to do?

- See if it is already in a database?
- Find similar sequences?
- Identify homologous sequences?
- Consider functional assignments and other annotation?
- Find primers in sequence?
- Find a short peptide?

UTILITY

- Relationships in evolution
- Identification of genes
- Assignment of possible functionality for genes or residues
- Possible structural understanding
- Aid in sequence assembly

What are you Comparing

- **Homologue**
  Sequences that share a common ancestor; may have similar function
- **Paralogue**
  Similar sequence within species, may have similar function
- **Orthologue**
  Same sequence separated by a speciation event, probably same function

ANALOG

Non-homologue proteins that have similar folding architecture, or similar functional sites, which are believed to have arisen through convergent evolution

Searching for Similarity

- BLAST
  - Search at NCBI and other servers (or locally)
  - Non-redundant set of databases, one DB at a time
  - Fast
  - Shows several similar regions

Searching for Similarity

- FASTA
  - Search against user-defined search sets, DB or subsections
  - Only the single most similar region is shown

The Word –Size Parameter

A word is any short sequence less than or equal to six letter

- Protein 1-2
- Nucleotide 1-6

High word Size

- Faster
- Less Sensitive
- More Selective

Evolution and Alignment

Evolutionary concepts enable the determination of similarity and homology

- Similarity is an observable quantity, such as %identity
- Homology is a conclusion drawn from the data that two genes share a common evolutionary history.

Evolution and Alignments (2)

- Genes are either homologous or not homologous.
- There is no degree of homology
- You can't tell what the ancestral sequence is simply because you have two or more homologues.

So, what IS an Alignment?

Evolution and Alignments (3)

- Alignments reflect the PROBABLE evolutionary history of two sequences
- Residues that align and are not identical represent substitutions
- Sequences without correspondence is aligned sequences are interpreted as indels and in an alignment are gaps.

Evolution and Alignment

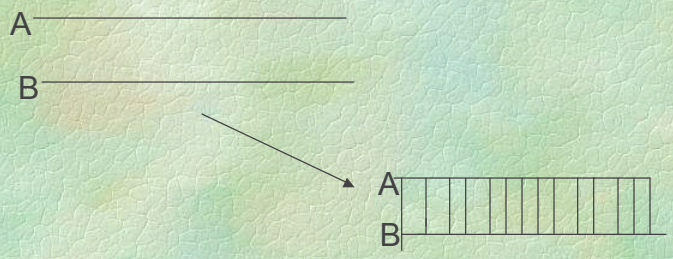
- Certain regions are more conserved than others, based on structure/function
- Certain regions may be conserved simply by history, not function
- This is true especially for closely related species.

Structure and Alignment

- If two proteins have more than 20-30% ID aligned, then the 3-D structures tend to be similar
- Overall folds are the same, details differ
- Form often follows function (Beware the BUT).
- So, sequence alignment is sometimes a 3-D alignment.

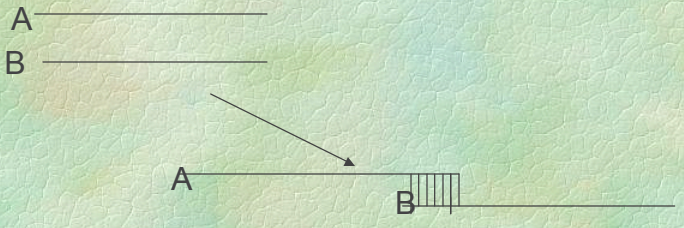
Global Alignment

Optimal alignment over the entire length



Local Alignment

Finds the highest coring alignment regardless of position and length



Needleman Wunsch Algorithm

- Global alignment:: every residue of the two sequences has to participate
- Guaranteed to calculate an Optimal similarity score
- Begin at the beginning of each sequence and go to the end.
- Cannot detect domains

Smith-Waterman Algorithm

- Optimal Local Alignment
- Guaranteed to find all significant matches to a given query
- Takes the query sequence versus every sequence in the database
- Can be used with arbitrary scoring systems
- **COMPUTATIONALLY EXPENSIVE!!!**

Scoring Matrices

- Relatively simple for DNA-gap penalties or mismatches-can be made to look at Pu/Py
- Protein matches look also at similarity (leu/ileu)

Searching for Similarity

- FASTA
  - Search against user-defined search sets, DB or subsections
  - Only the single most similar region is shown

Nucleotide Uncertainties

Code	Meaning (Base)	Code	Meaning (Base)
A	adenosine (A)	M	amino (A or C)
C	cytidine (C)	S	strong (G or C)
G	guanine (G)	W	weak (A or T)
T	thymidine (T)	B	not A (G or T or C)
U	uridine (U)	D	not C G or A or T)
R	purine (G or A)	H	not G (A or C or T)
Y	pyrimidine (T or C)	V	not T (G or C or A)
K	keto (G or T)	N	any base (A or G or C or T)
-	gap(s) (none)		

Protein Scoring Matrices

- Chemical similarity: 210 pairs of aa
- Nearness in Genetic Code
- Chemical similarity, e.g., hydrophobicity
- Observed Substitution Schemes

AA Substitution Matrices

Rationale:

Certain amino acid substitutions commonly occur in related proteins (sometimes from different species). These provide the basis for amino acid substitution matrices, essentially a symbol comparison table.

More on Matrices

- A substitution matrix specifies a set of scores s_{ij} for replacing amino acid I by amino acid j.
- PAM: Percent Accepted Mutations
- BLOSUM Blocks Amino Acid Substitution Matrices

Amino Acid Symbols

- A Ala alanine
- B Asx Aspartic or asparaine
- C Cys Cysteine
- D Aspartic acid
- E Glu Glutamic acid
- F Phe Phenylalanine
- G Gly Glycine
- H His Histidine
- I Ile Isoleucine
- K Lys Lysine
- L Leu Leucine
- M Met Methionine
- N Asn Asparagine
- P Pro Proline
- Q Gln Glutamine
- R Arg Argine
- S Ser Serine
- T Thr Threonine
- U Sec Selenocysteine
- V Val Valine
- W Trp Tryptophan
- X Xaa Unknown or other aa
- Y Tyr Tyrsoine
- Z Glx Glutamic or glutamine

Observed AA Substitution Matrices

- PAM
- BLOSUM

PAM

- Log Odds scores are used
- The score of each pair s(a,b) is defined as the log of the likelihood ratio of the transition probability M_{ab} (Mutation) versus the probability of a random occurrence of the amino acid b in the second sequence.

$$s(a,b)=\log M_{ab}/P_b$$

PAM: Point Accepted Mutation

- DAYHOFF et al.
- Observed residue replacement in related proteins
- GLOBAL alignment, closely related
- A model of molecular evolution

1 PAM = average change in 1% of all amino acid possibilities(1% divergence)

- Other PAM matrices extrapolated from PAM1.

PAM continued

- TIME is NOT correlated with PAM
- Number of the matrix refers to evolutionary distance

Means different families of proteins evolve at different rates

PAM250

Table 2: The PAM250 matrix – an example of a matrix derived from observed substitutions

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	2	0	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	0
B	0	2	-4	3	2	-5	0	1	-2	1	-3	-2	2	-1	1	-1	0	0	-2	-5	-3	2
C	-2	-4	12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	-5
D	0	3	-5	4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	3
E	0	2	-5	3	4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4	3
F	-4	-5	-4	-6	-5	9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7	-5
G	1	0	-3	1	0	-5	5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5	-1
H	-1	1	-3	1	1	-2	-2	6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0	2
I	-1	-2	-2	-2	-2	1	-3	-2	5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1	-2
K	-1	1	-5	0	0	-5	-2	0	-2	5	-3	0	1	-1	1	3	0	0	-2	-3	-4	0
L	-2	-3	-6	-4	-3	2	-4	-2	2	-3	6	4	-3	-3	-2	-3	-3	-2	2	-2	-1	-3
M	-1	-2	-5	-3	-2	0	-3	-2	2	0	4	6	-2	-2	-1	0	-2	-1	2	-4	-2	-2
N	0	2	-4	2	1	-4	0	2	-2	1	-3	-2	2	-1	1	0	1	0	-2	-4	-2	1
P	1	-1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6	0	0	1	0	-1	-6	-5	0
Q	0	1	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4	1	-1	-1	-2	-5	-4	3
R	-2	-1	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6	0	-1	-2	2	-4	0
S	1	0	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2	1	-1	-2	-3	0
T	1	0	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3	0	-5	-3	-1
V	0	-2	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4	-6	-2	-2
W	-6	-5	-8	-7	-7	0	-7	-3	5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	0	-6
Y	-3	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10	-4
Z	0	2	-5	3	3	-5	-1	2	-2	0	-3	-2	1	0	3	0	0	-1	-2	-6	-4	3

The non-standard amino acid B is used to refer to (D or N); Z refers to (E or Q).

BLOSUM

- Block Substitution Matrix
- Henikoff and Henikoff, PNAS, 1992
- Number following indicates per cent identity within set, BLOSUM62=62% id
- Finds short, highly similar sequences (no gaps)

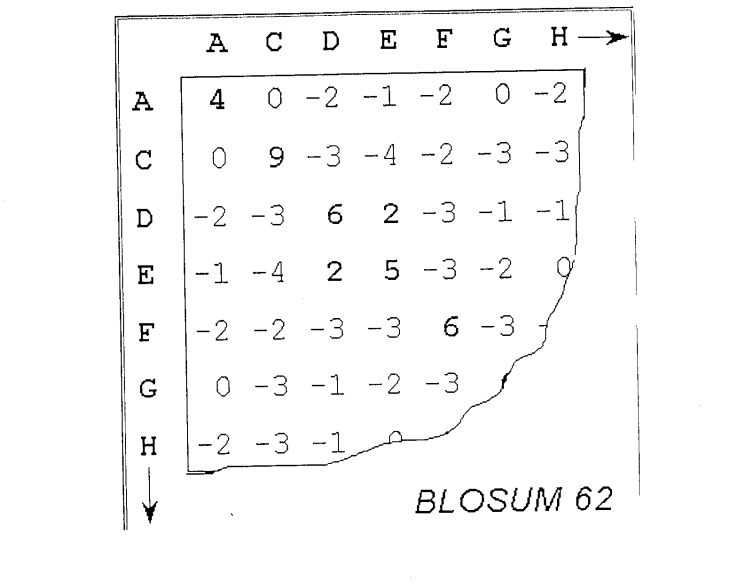
BLOSUM

- Matrices are directly calculated, based on observed alignments
- Greater numbers are lesser distances
- Usually best for local similarity searches
- BLOSUM62= DEFAULT FOR BLAST. If a distant relative, think about another matrix.

BLOSUM SCORING RULES

- Zero score means the frequencies of the pair in the database is that expected by chance
- A positive score means more frequent than chance
- Negative score means the pair is found less frequently than chance.

Blosum62

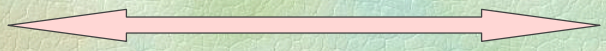


	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H ↓	-2	-3	-1	0			

BLOSUM 62

BLOSUM80 BLOSUM62 BLOSUM45

PAM1 PAM120 PAM250



less divergent more divergent

BLAST-Basic Local Alignment Sequence Tool

- Objective: find all local regions of similarity distinguishable from random
- Only local alignments permitted,
- Gaps permitted in version 2
- Statistically sound (Karlin and Altschul), but no guantee of optimality

BLAST: Three Step Algorithm

- Compile a list of high scoring words of length w (w=4 for proteins, 12 for nucleic acids)
- Scan for word hits of score greater than threshold, T
- Extend word hit in both directions to find High Scoring Pairs with scores greater than S

Other BLAST Programs

- BLASTN: nucleic acid query to NA database
- BLASTP: Protein query to Protein database
- BLASTX: Translated nucleic acid query to Protein database
- TBLASTN: Protein query against (translated) nucleic acid database
- TBLASTX: Translated nucleic acid against translated nucleic acid database

OTHER BLAST VARIATIONS

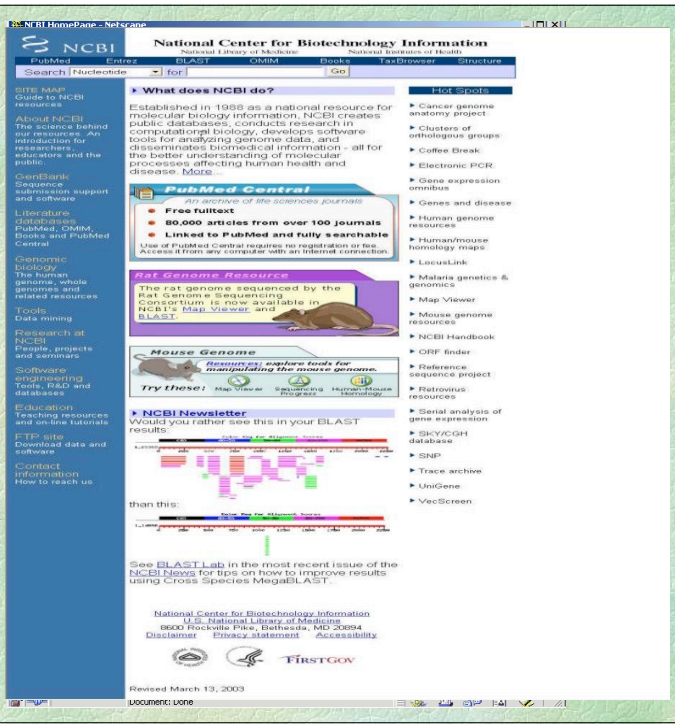
- PSI-BLAST- Position Specific Iterated BLAST-use gapped BLAST, generate a Profile from multiple iterations used instead of the input and Distance Matrix
- MEGABLAST specifically designed to efficiently find long alignments between very similar sequences; the best tool to use to find the identical match to your query sequence.

Limitations to BLAST

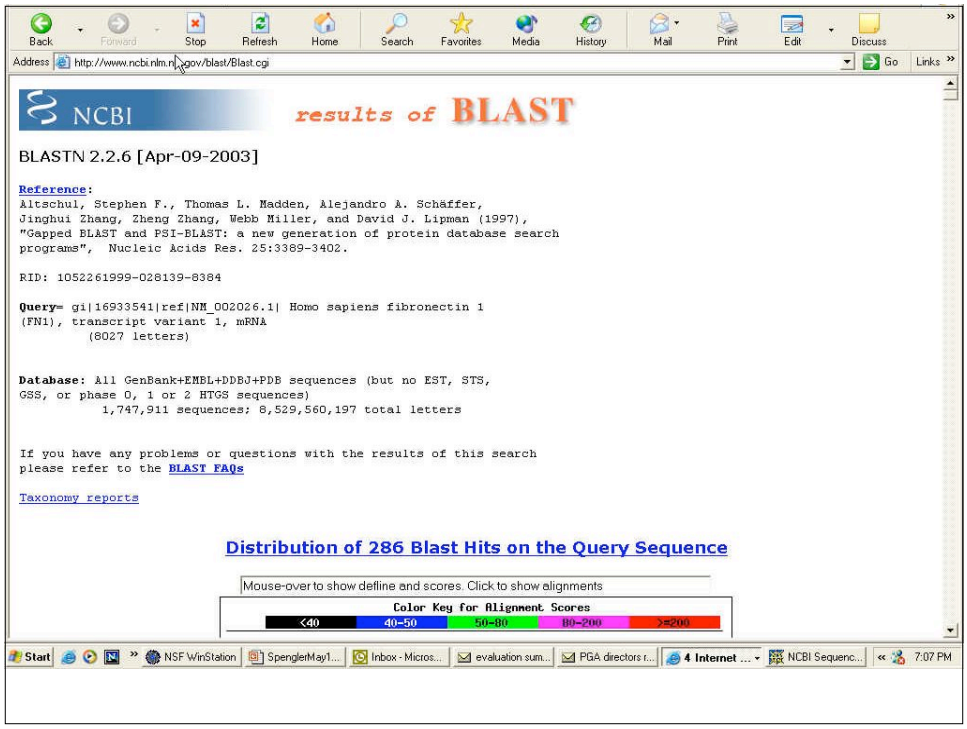
- Needs islands of strong homology
- Limits on the combination of scoring and penalty values
- The variants (blastx, tblastn, tblastx) use 6-frame translation-miss sequences with frameshifts)
- Finds and reports ONLY local alignments

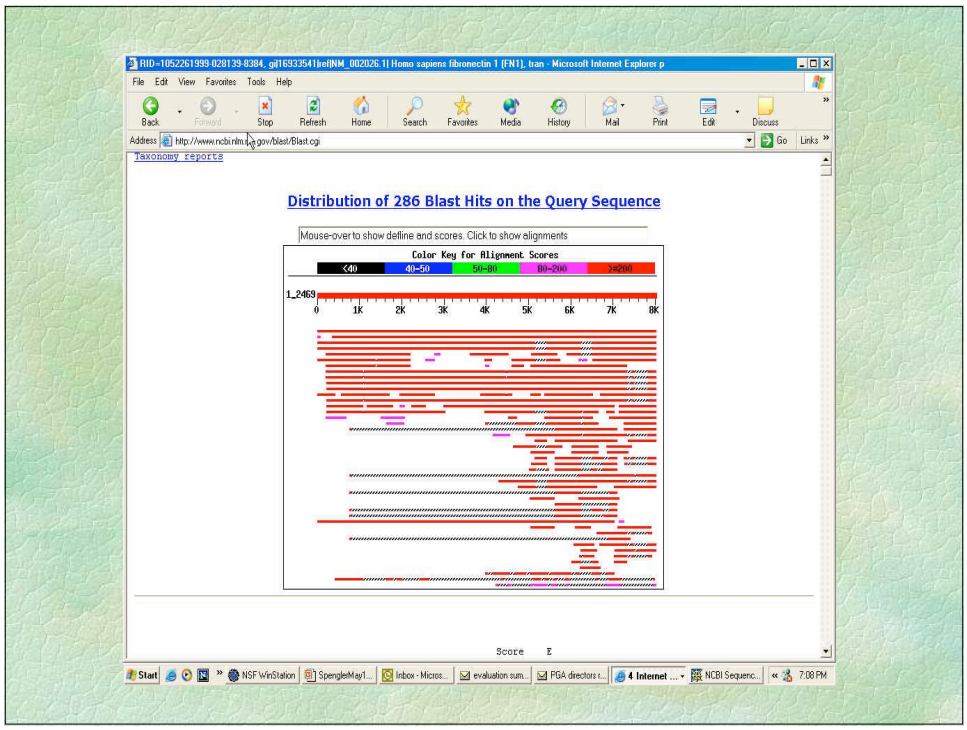
A WALK THROUGH BLAST

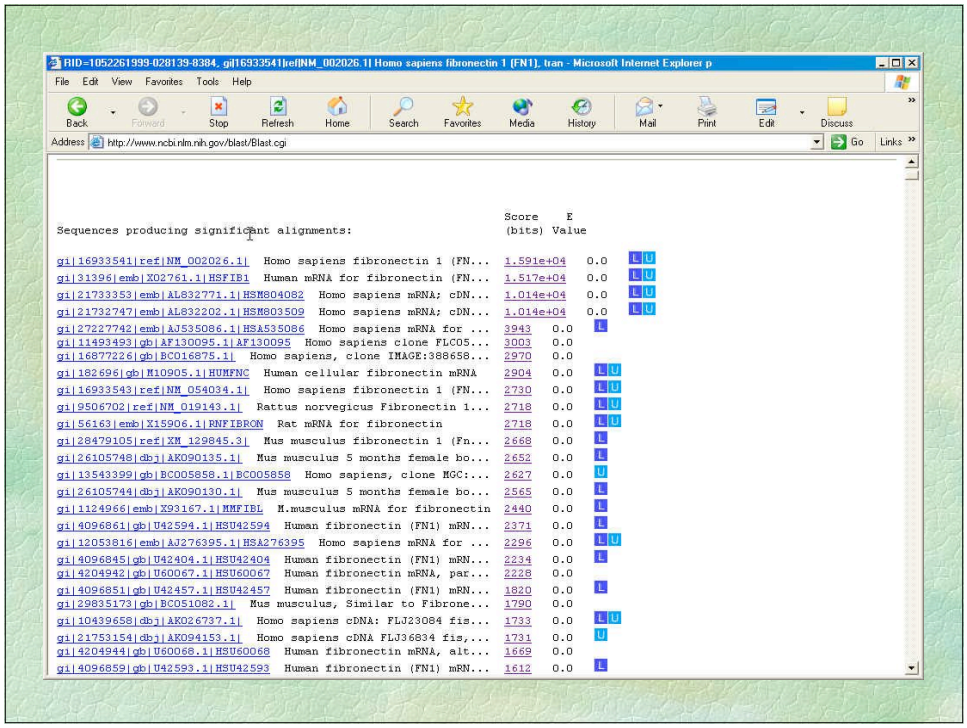
NCBI home

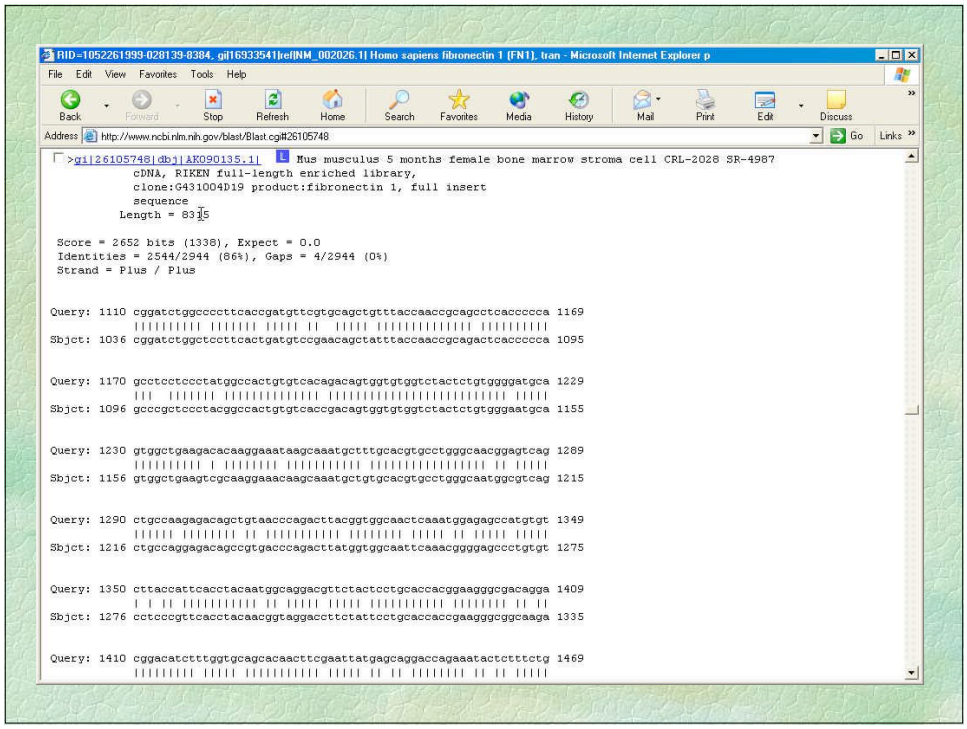


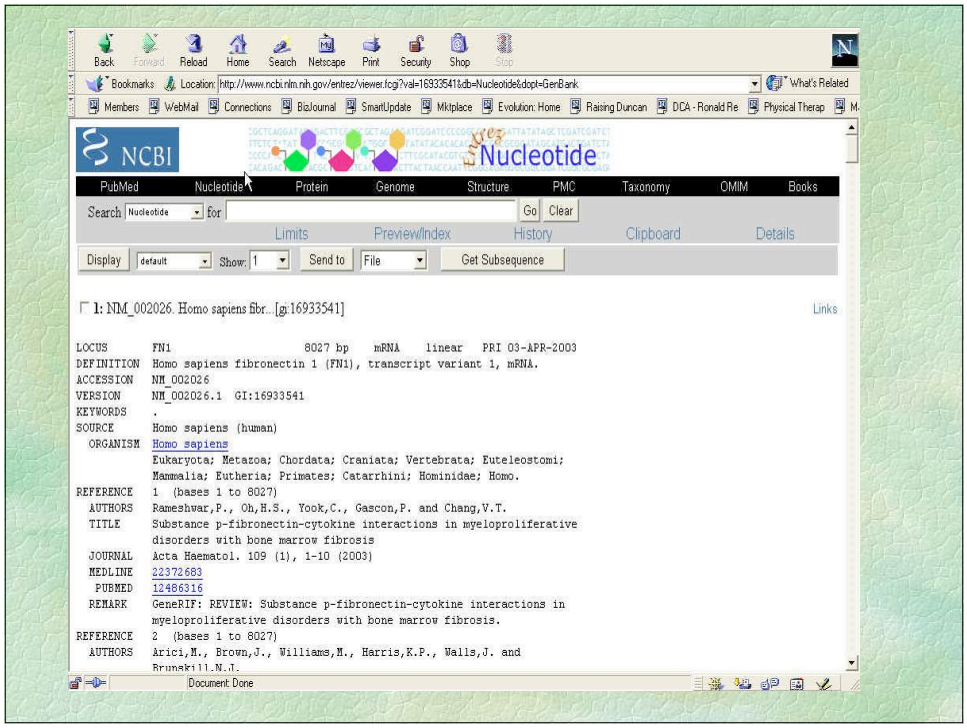

NCBI
formatting BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = (8027 letters)
Your search was limited by an Entrez query: >(none)
The request ID is 1052165600-01386-27393
Format!
or
Reset all
The results are estimated to be ready in 21 minutes but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your re
via the form below and press "FORMAT!" again. You may also request results of a different search by entering an
other valid request ID to see other recent jobs.

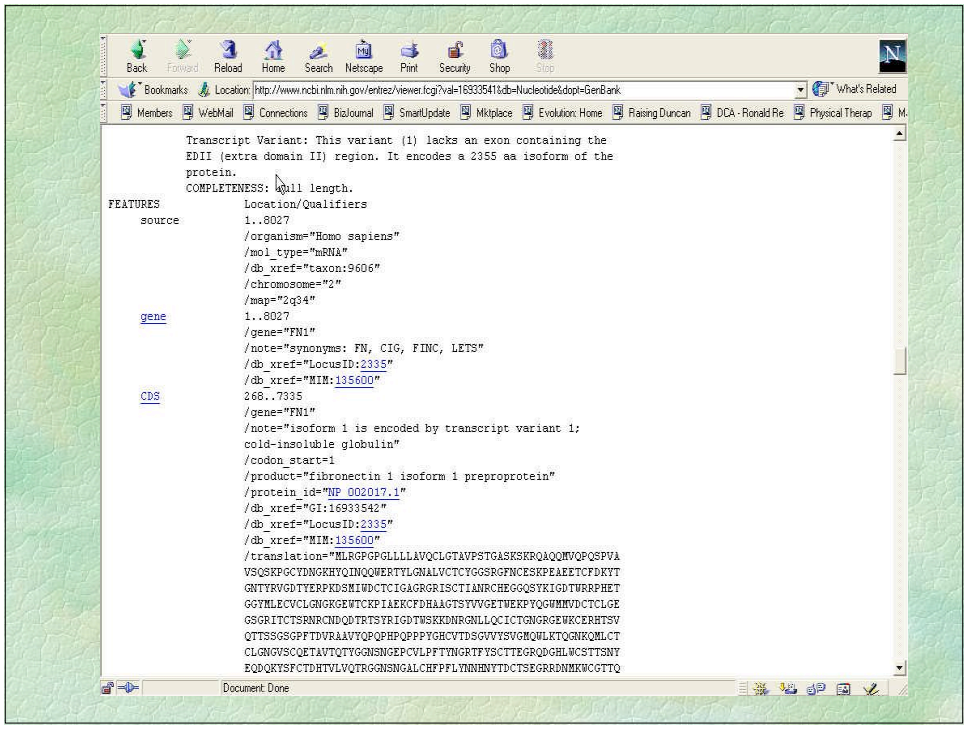

Back
Forward
Stop
Refresh
Home
Search
Favorites
Media
History
Mail
Print
Edit
Discuss
Address
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
Links
NCBI
results of BLAST
BLASTN 2.2.6 [Apr-09-2003]
Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.
RID: 1052261999-028139-8384
Query= gi|16933541|ref|NM_002026.1| Homo sapiens fibronectin 1
(FN1), transcript variant 1, mRNA
(8027 letters)
Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS, or phase 0, 1 or 2 HTGS sequences)
1,747,911 sequences; 8,529,560,197 total letters
If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports
Distribution of 286 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
Start
NSF WinStation
SpenglerMay1...
Inbox - Micros...
evaluation sum...
PGA directors r...
4 Internet ...
NCBI Sequenc...
7:07 PM

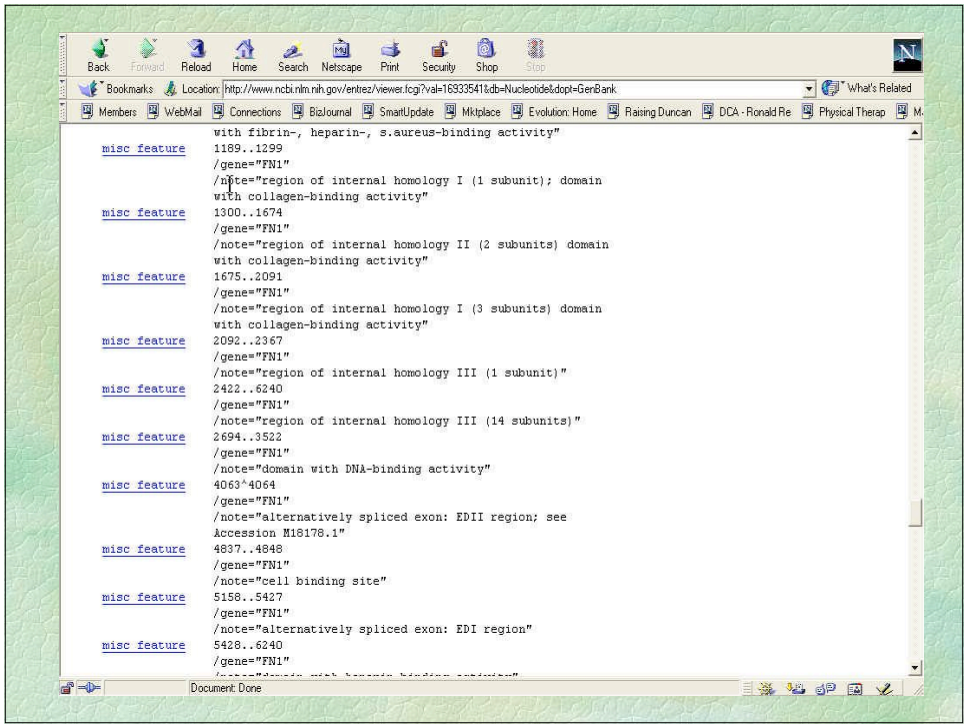

RID=1052261999-028139-8384, gi|16933541|ref|NM_002026.1| Homo sapiens fibronectin 1 (FN1), tran - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go Links
Taxonomy reports
Distribution of 286 Blast Hits on the Query Sequence
Mouse-over to show define and scores. Click to show alignments
Color Key for Alignment Scores
<40 40-50 50-80 80-200 >=200
1_2469
0 1K 2K 3K 4K 5K 6K 7K 8K
Score E
Start NSF WinStation SpengleMay1... Inbox - Micros... evaluation sum... PGA directors r... 4 Internet ... NCBI Sequenc... 7:08 PM

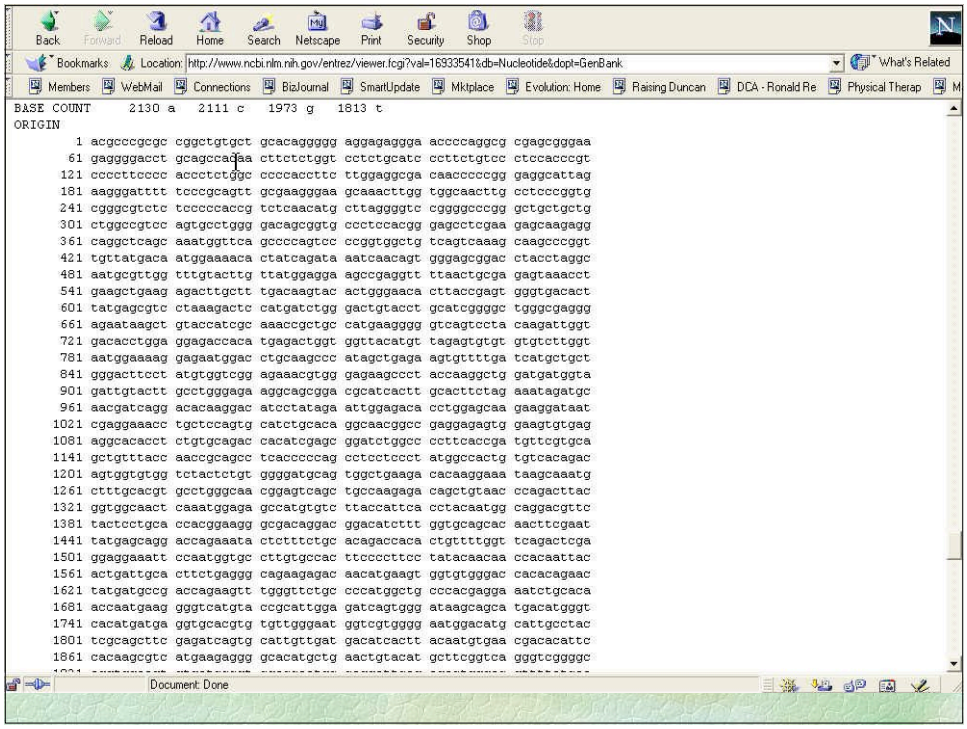

RID=1052261999-028139-8384, gi|16933541|ref|NM_002026.1| Homo sapiens fibronectin 1 (FN1), tran - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go Links
Sequences producing significant alignments: Score (bits) E Value
gi|16933541|ref|NM_002026.1| Homo sapiens fibronectin 1 (FN... 1.591e+04 0.0 L U
gi|31396|emb|X02761.1|HSFIB1 Human mRNA for fibronectin (FN... 1.517e+04 0.0 L U
gi|21733353|emb|AL832771.1|HSM804082 Homo sapiens mRNA; cDN... 1.014e+04 0.0 L U
gi|21732747|emb|AL832202.1|HSM803509 Homo sapiens mRNA; cDN... 1.014e+04 0.0 L U
gi|27227742|emb|AJ535086.1|HSA535086 Homo sapiens mRNA for ... 3943 0.0 L
gi|11493493|gb|AF130095.1|AF130095 Homo sapiens clone FLC05... 3003 0.0
gi|16877226|gb|BC016875.1| Homo sapiens, clone IMAGE:388658... 2970 0.0
gi|182696|gb|M10905.1|HUMFNC Human cellular fibronectin mRNA 2904 0.0 L U
gi|16933543|ref|NM_054034.1| Homo sapiens fibronectin 1 (FN... 2730 0.0 L U
gi|9506702|ref|NM_019143.1| Rattus norvegicus Fibronectin 1... 2718 0.0 L U
gi|56163|emb|X15906.1|RNFIBRON Rat mRNA for fibronectin 2718 0.0 L U
gi|28479105|ref|XM_129845.3| Mus musculus fibronectin 1 (Fn... 2668 0.0 L
gi|26105748|dbj|AK090135.1| Mus musculus 5 months female bo... 2652 0.0 L
gi|13543399|gb|BC005858.1|BC005858 Homo sapiens, clone MGC:... 2627 0.0 U
gi|26105744|dbj|AK090130.1| Mus musculus 5 months female bo... 2565 0.0 L
gi|1124966|emb|X93167.1|MMFIBL M.musculus mRNA for fibronectin 2440 0.0 L
gi|4096861|gb|U42594.1|HSU42594 Human fibronectin (FN1) mRN... 2371 0.0 L
gi|12053816|emb|AJ276395.1|HSA276395 Homo sapiens mRNA for ... 2296 0.0 L U
gi|4096845|gb|U42404.1|HSU42404 Human fibronectin (FN1) mRN... 2234 0.0 L
gi|4204942|gb|U60067.1|HSU60067 Human fibronectin mRNA, par... 2228 0.0
gi|4096851|gb|U42457.1|HSU42457 Human fibronectin (FN1) mRN... 1820 0.0 L
gi|29835173|gb|BC051082.1| Mus musculus, Similar to Fibrone... 1790 0.0
gi|10439658|dbj|AK026737.1| Homo sapiens cDNA: FLJ23084 fis... 1733 0.0 L U
gi|21753154|dbj|AK094153.1| Homo sapiens cDNA FLJ36834 fis,... 1731 0.0 U
gi|4204944|gb|U60068.1|HSU60068 Human fibronectin mRNA, alt... 1669 0.0
gi|4096859|gb|U42593.1|HSU42593 Human fibronectin (FN1) mRN... 1612 0.0 L


RID=1052261999-028139-8384, gi|16933541|ref|NM_002026.1| Homo sapiens fibronectin 1 (FN1), tran - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#26105748
>gi|26105748|dbj|AK090135.1| Mus musculus 5 months female bone marrow stroma cell CRL-2028 SR-4987
cDNA, RIKEN full-length enriched library,
clone:G431004D19 product:fibronectin 1, full insert
sequence
Length = 8315
Score = 2652 bits (1338), Expect = 0.0
Identities = 2544/2944 (86%), Gaps = 4/2944 (0%)
Strand = Plus / Plus
Query: 1110 cggatctggccccttcaccgatgttcgtgcagctgtttaccaaccgcagcctcaccccca 1169
Sbjct: 1036 cggatctggctccttcactgatgtccgaacagctatttaccaaccgcagactcaccccca 1095
Query: 1170 gcctcctccctatggccactgtgtcacagacagtggtgtggtctactctgtggggatgca 1229
Sbjct: 1096 gcccgctccctacggccactgtgtcaccgacagtggtgtggtctactctgtgggaatgca 1155
Query: 1230 gtggctgaagacacaaggaaataagcaaatgctttgcacgtgcctgggcaacggagtcag 1289
Sbjct: 1156 gtggctgaagtcgcaaggaaacaagcaaatgctgtgcacgtgcctgggcaatggcgtcag 1215
Query: 1290 ctgccaagagacagctgtaacccagacttacggtggcaactcaaatggagagccatgtgt 1349
Sbjct: 1216 ctgccaggagacagccgtgacccagacttatggtggcaattcaaacggggagccctgtgt 1275
Query: 1350 cttaccattcacctacaatggcaggacgttctactcctgcaccacggaagggcgacagga 1409
Sbjct: 1276 cctcccgttcacctacaacggtaggaccttctattcctgcaccaccgaagggcggcaaga 1335
Query: 1410 cggacatctttggtgcagcacaacttcgaattatgagcaggaccagaaatactcttttctg 1469


Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=16933541&db=Nucleotide&dopt=GenBank
What's Related
Members WebMail Connections BizJournal SmartUpdate Mktplace Evolution: Home Raising Duncan DCA - Ronald Re Physical Therap
NCBI
Entrez Nucleotide
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search Nucleotide for Go Clear
Limits Preview/Index History Clipboard Details
Display default Show: 1 Send to File Get Subsequence
1: NM_002026. Homo sapiens fibr... [gi:16933541]
Links
LOCUS FN1 8027 bp mRNA linear PRI 03-APR-2003
DEFINITION Homo sapiens fibronectin 1 (FN1), transcript variant 1, mRNA.
ACCESSION NM_002026
VERSION NM_002026.1 GI:16933541
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 8027)
AUTHORS Rameshwar,P., Oh,H.S., Yook,C., Gascon,P. and Chang,V.T.
TITLE Substance p-fibronectin-cytokine interactions in myeloproliferative
disorders with bone marrow fibrosis
JOURNAL Acta Haematol. 109 (1), 1-10 (2003)
MEDLINE 22372683
PUBMED 12486316
REMARK GeneRIF: REVIEW: Substance p-fibronectin-cytokine interactions in
myeloproliferative disorders with bone marrow fibrosis.
REFERENCE 2 (bases 1 to 8027)
AUTHORS Arici,M., Brown,J., Williams,M., Harris,K.P., Walls,J. and
Brunskill,N.J.
Document: Done


Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=16933541&db=Nucleotide&dopt=GenBank
What's Related
Members WebMail Connections BizJournal SmartUpdate MktIplace Evolution: Home Raising Duncan DCA - Ronald Re Physical Therap
Transcript Variant: This variant (1) lacks an exon containing the
EDII (extra domain II) region. It encodes a 2355 aa isoform of the
protein.
COMPLETENESS: full length.
FEATURES Location/Qualifiers
source 1..8027
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="taxon:9606"
/chromosome="2"
/map="2q34"
gene 1..8027
/gene="FN1"
/note="synonyms: FN, CIG, FINC, LETS"
/db_xref="LocusID:2335"
/db_xref="MIM:135600"
CDS 268..7335
/gene="FN1"
/note="isoform 1 is encoded by transcript variant 1;
cold-insoluble globulin"
/codon_start=1
/product="fibronectin 1 isoform 1 preproprotein"
/protein_id="NP_002017.1"
/db_xref="GI:16933542"
/db_xref="LocusID:2335"
/db_xref="MIM:135600"
/translation="MLRGPGPGLLLLAVQCLGTAVPSTGASKSKRQAQQMVQPQSPVA
VSQSKPGCYDNGKHYQINQQWERTYLGNALVCTCYGGSRGFNCESKPEAEETCFDKYT
GNTYRVGDTYERPKDSMIWDCTCIGAGRGRISCTIANRCHEGGQSYKIGDTWRRPHET
GGYMLECVCLGNGKGEWTCKPIAEKCFDHAAGTSYVVGETWEKPYQGWMMVDCTCLGE
GSGRITCTSRNRCNDQDTRTSYRIGDTWSKKDNRGNLLQCICTGNGRGEWKCERHTSV
QTTSSGSGPFTDVRAAVYQPQPHPQPPPYGHCVTDSGVVYSVGMQWLKTQGNKQMLCT
CLGNGVSCQETAVTQTYGGNSNGEPCVLPFTYNGRTFYSCTTEGRQDGHLWCSTTSNY
EQDQKYSFCTDHTVLVQTRGGNSNGALCHFPFLYNNHNVTDCTSEGRRDNMKWCGTTQ
Document: Done

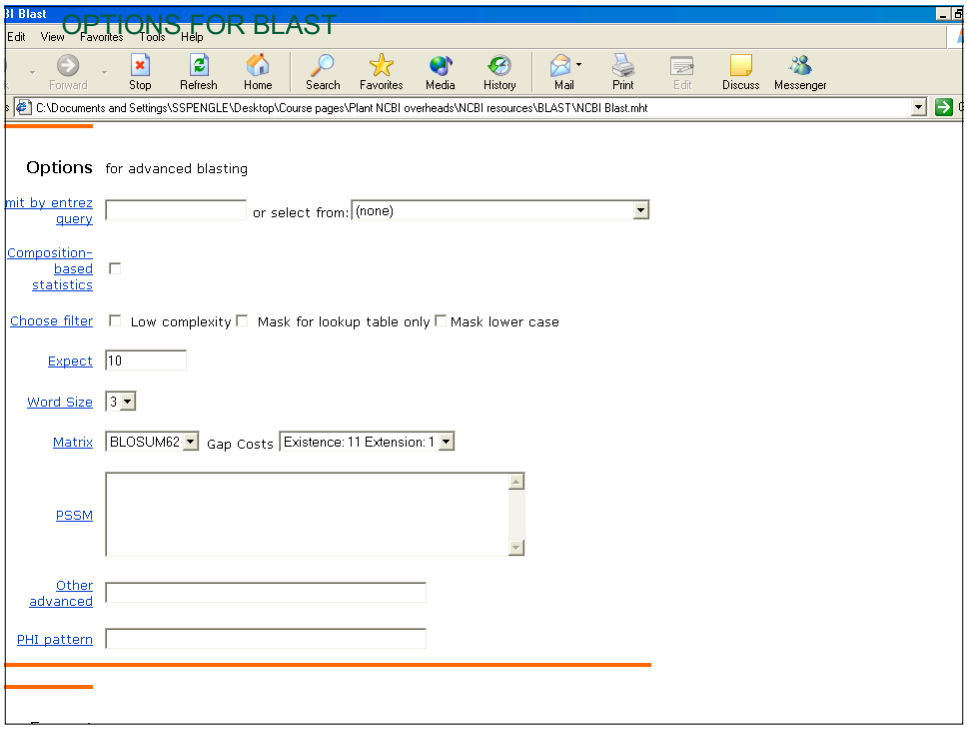

Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=16933541&db=Nucleotide&dopt=GenBank
What's Related
Members WebMail Connections BizJournal SmartUpdate MktIplace Evolution: Home Raising Duncan DCA - Ronald Re Physical Therap
with fibrin-, heparin-, s.aureus-binding activity"
misc feature 1189..1299
/gene="FN1"
/note="region of internal homology I (1 subunit); domain
with collagen-binding activity"
misc feature 1300..1674
/gene="FN1"
/note="region of internal homology II (2 subunits) domain
with collagen-binding activity"
misc feature 1675..2091
/gene="FN1"
/note="region of internal homology I (3 subunits) domain
with collagen-binding activity"
misc feature 2092..2367
/gene="FN1"
/note="region of internal homology III (1 subunit)"
misc feature 2422..6240
/gene="FN1"
/note="region of internal homology III (14 subunits)"
misc feature 2694..3522
/gene="FN1"
/note="domain with DNA-binding activity"
misc feature 4063^4064
/gene="FN1"
/note="alternatively spliced exon: EDII region; see
Accession M18178.1"
misc feature 4837..4848
/gene="FN1"
/note="cell binding site"
misc feature 5158..5427
/gene="FN1"
/note="alternatively spliced exon: EDI region"
misc feature 5428..6240
/gene="FN1"
Document: Done



What happens with changes?

- Word size
- Substitution Matrix
- Expected Value
- Using protein or translated sequence rather than nucleotide

OPTIONS FOR BLAST



The Word–Size Parameter

A word is any short sequence less than or equal to

- Protein 1-2
- Nucleotide 1-12

Higher word Size

- Faster
- Less Sensitive
- More Selective

Expected Value

- Number of hits with a score better than "S" expected by chance in a DB of a given size
- Default value of threshold is 10.
- For a short sequence, increase the E value (say to 1000)

BLAST RULES OF THUMB

- For short amino acid sequences (20-40), 50% identity happens by chance
- If A and B are homologous, and B and C are homologous, then A and C are, even if you can't see it.
- You can get similarity in the absence of homology for low complexity, transmembrane and coiled-coil regions. These have to be eliminated by you, but you MAY want them.

TAKE-HOMES

- Use an up-to-date database; repeat often
- Choose a fast algorithm
- Use the most recent version
- Work at the protein level--for a small amount of evolutionary change, DNA sequence contains less information about homology
- Respect your own ***intuition***

BLAST Significance

- If you change scoring systems, you can still compare search results if you normalize the score.

  S'=(lambdaS-lnK)/ln2. Lambda and K are associated with the scoring system.

  S', with a given E, is significant if it is greater than log N/E, N the size of the search space.

Searching for Similarity

- FASTA
  - Search against user-defined search sets, DB or subsections
  - Only the single most similar region is shown

FASTA: WHY USE IT?

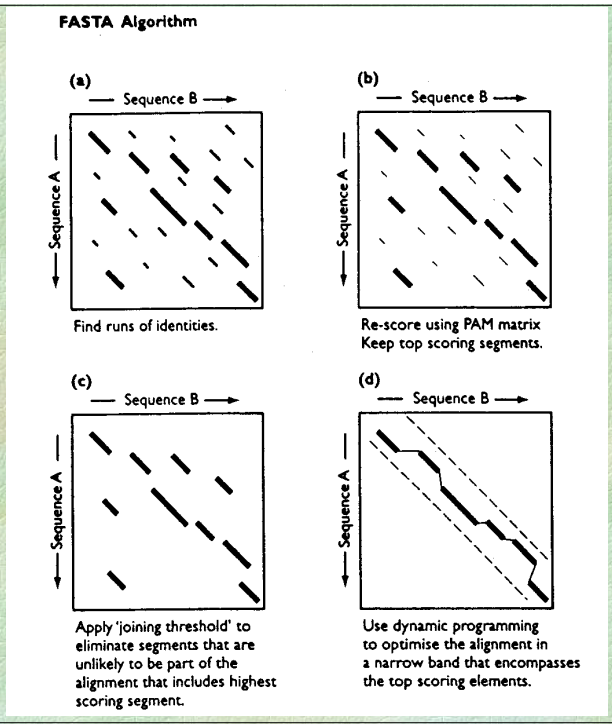
- Allow alignments to shift frames

FASTA: FAST Alignment

- http://alpha10.bioch.virginia.edu/fasta/
- http://www2.ebi.ac.uk/fasta3
- http://workbench.sdsc.edu

- Rapid Global alignment

- Not a strong mathematical basis



LALIGN

- Essentially a FASTA derivative for local alignments
- Compares two proteins to identify regions of similarity
- Will report several sequence alignments within a given sequence
- Works for internal repeats that are missed by FASTA because of gaps.

SITEs for LALIGN

- http://fasta.bioch.virginia.edu/fasta/lalign.htm
- http://xylian.igh.cnrs.fr/bin/lalign-guess.cgi
- http://biowb.sdsc.edu (registration necessary but painless)
- PALIGN http://fasta.bioch.virginia.edu/fasta/palign.htm (plots a graph of the areas of alignment)

Gene Ontologies GO

A gene ontology is a controlled vocabulary used to describe the biology of a gene product in any organism, designed to allow both attribution and querying at different levels of granularity , facilitating queries across participating databases.

A step toward unifying biological databases but not sufficient.

http://www.geneontlogy.org

Components of GO

A gene product is a physical thing (protein, RNA, can have small molecules associated to make a gene product group.

Attributes of Gene Products

- **Molecular Function**-what something does
- **Biological process**-a biological objective, like growth or pyrimidine metabolism
- **Cellular Component-**part of a cell, ER, nucleus etc.

Ontology Representations

- A network, a directed acyclic graph (DAG), in which terms have multiple parents and multiple relationships to parents.
- Relationships connecting terms include is-a, part-of,

Yeast, Fly, Mouse, Arabidopsis, Worm,

EVIDENCE CODES

- IC Inferred by Curator
- IDA Inferred by Direct Assay
- IEA Inferred by Electronic Annotation
- IEP Inferred from expression patter
- IGI Inferred from genetic interaction
- IMP Inferred from mutant phenotype
- IPI Inferred from physical interaction
- ISS Inferred from sequence or structure similarity
- NAS Non-traceable author statement
- ND No biological data available
- TAS Traceable author statement
- NR Not recorded

Evidence relationships

TAS/IDA

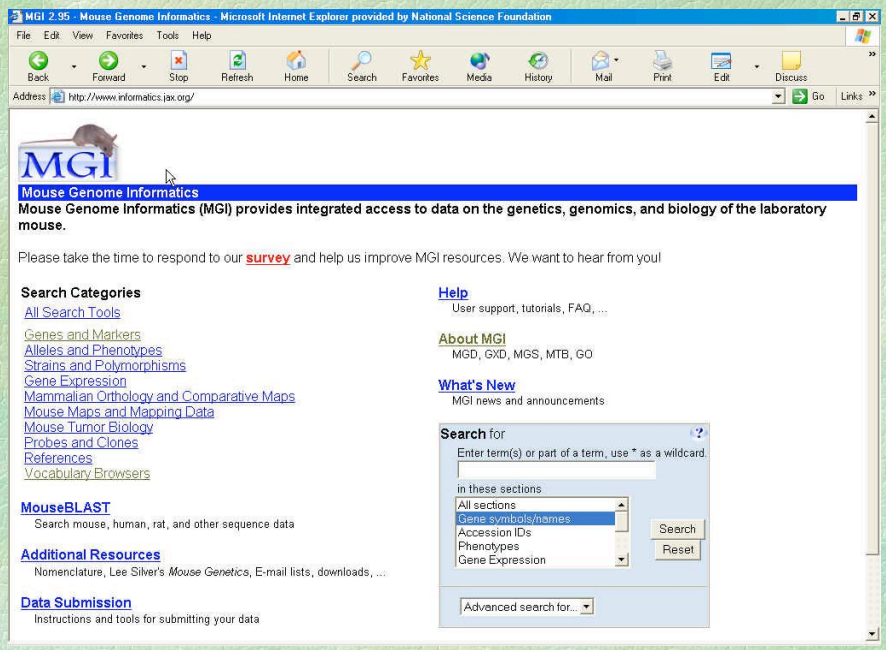
IMP/IGI/IPI

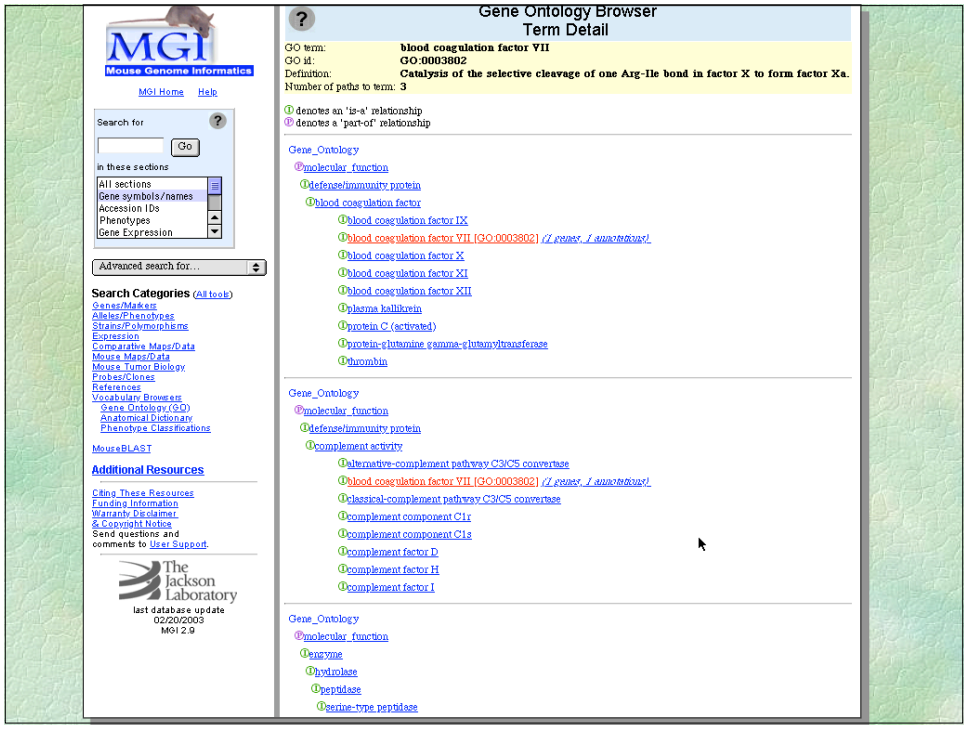
ISS/IEP

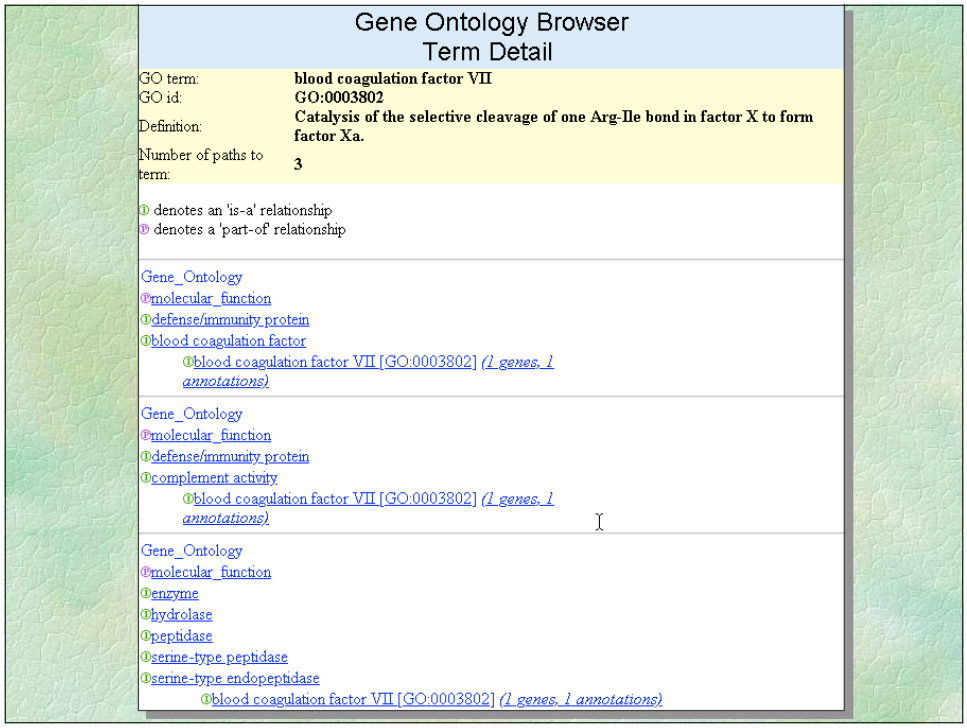
NAS

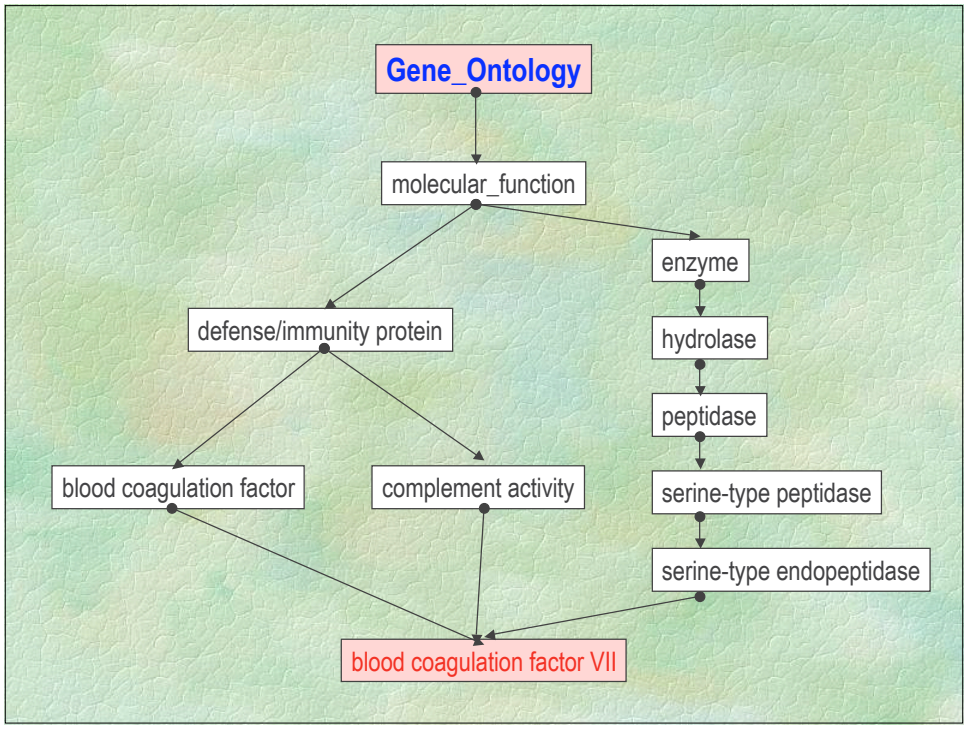
IEA

Not a rigid hierarchy.




MGI
Mouse Genome Informatics
MGI Home Help
Search for
Go
in these sections
All sections
Gene symbols/names
Accession IDs
Phenotypes
Gene Expression
Advanced search for...
Search Categories (All tools)
Genes/Markers
Alleles/Phenotypes
Strains/Polymorphisms
Expression
Comparative Maps/Data
Mouse Maps/Data
Mouse Tumor Biology
Probes/Clones
References
Vocabulary Browsers
Gene Ontology (GO)
Anatomical Dictionary
Phenotype Classifications
MouseBLAST
Additional Resources
Citing These Resources
Funding Information
Warranty Disclaimer
& Copyright Notice
Send questions and comments to User Support.
The Jackson Laboratory
last database update
02/20/2003
MGI 2.9
Gene Ontology Browser
Term Detail
GO term: blood coagulation factor VII
GO id: GO:0003802
Definition: Catalysis of the selective cleavage of one Arg-Ile bond in factor X to form factor Xa.
Number of paths to term: 3
denotes an 'is-a' relationship
denotes a 'part-of' relationship
Gene_Ontology
molecular_function
defense/immunity protein
blood coagulation factor
blood coagulation factor IX
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)
blood coagulation factor X
blood coagulation factor XI
blood coagulation factor XII
plasma kallikrein
protein C (activated)
protein-glutamine gamma-glutamyltransferase
thrombin
Gene_Ontology
molecular_function
defense/immunity protein
complement activity
alternative-complement pathway C3/C5 convertase
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)
classical-complement pathway C3/C5 convertase
complement component C1r
complement component C1s
complement factor D
complement factor H
complement factor I
Gene_Ontology
molecular_function
enzyme
hydrolase
peptidase
serine-type peptidase

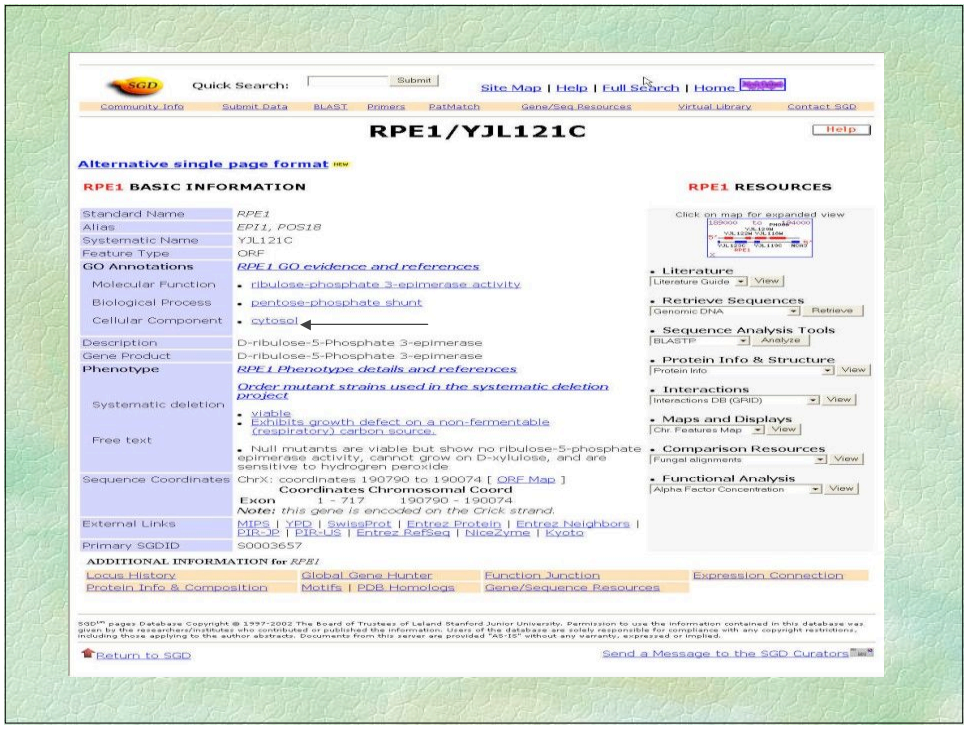

Gene Ontology Browser
Term Detail
GO term: blood coagulation factor VII
GO id: GO:0003802
Definition: Catalysis of the selective cleavage of one Arg-Ile bond in factor X to form factor Xa.
Number of paths to term: 3
denotes an 'is-a' relationship
denotes a 'part-of' relationship
Gene_Ontology
molecular_function
defense/immunity protein
blood coagulation factor
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)
Gene_Ontology
molecular_function
defense/immunity protein
complement activity
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)
Gene_Ontology
molecular_function
enzyme
hydrolase
peptidase
serine-type peptidase
serine-type endopeptidase
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)

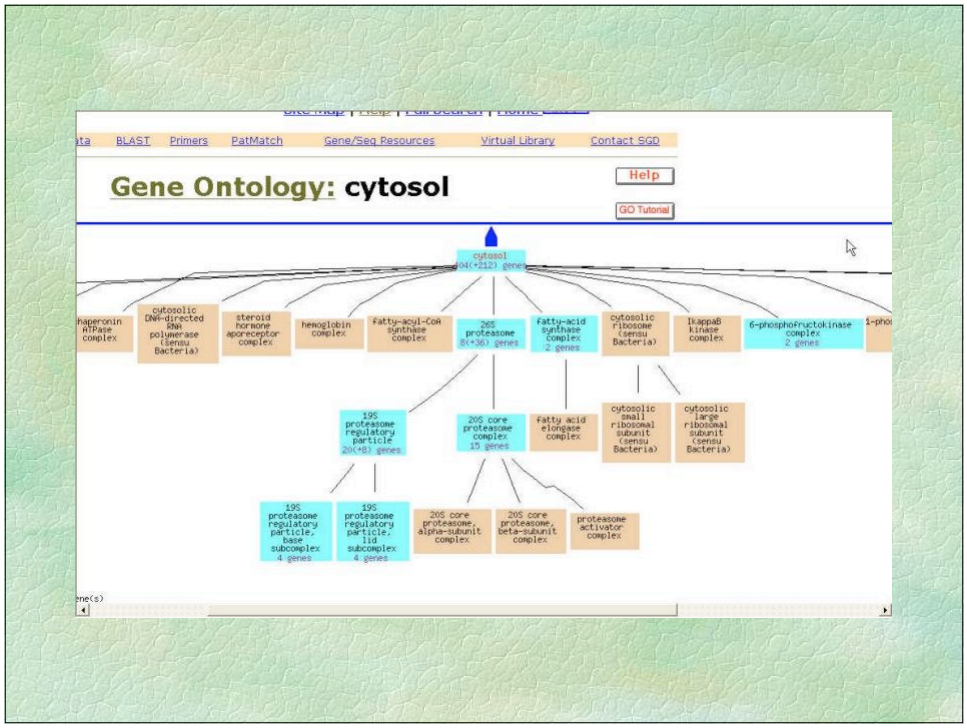


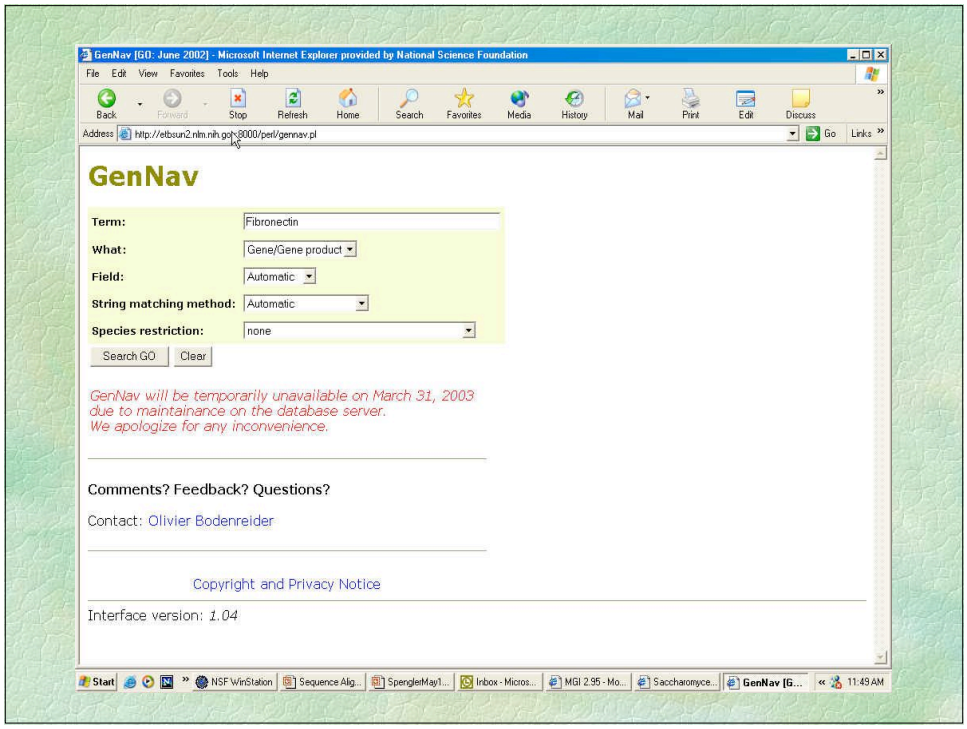
GO Browsers	
AmiGO from BDGP	• With AmiGO, you can search for a GO term and view all gene products annotated to it, or search for a gene product and view all its associations. You can also browse the ontologies to view relationships between terms as well as the number of gene products annotated to a given term. AmiGO accesses the GO mySQL database (see below); the browser and documentation are available from http://www.godatabase.org/dev/
MGI GO Browser	• With the MGI GO Browser, you can search for a GO term and view all mouse genes annotated to the term or any subterms. You can also browse the ontologies to view relationships between terms, term definitions, as well as the number of mouse genes annotated to a given term and its subterms. The MGI GO browser directly accesses the GO in the MGI database where mouse gene annotations, are updated nightly. The version of the GO used is obtained nightly from the GO ftp site.
QuickGO at EBI	• With QuickGO, a GO browser integrated into InterPro at the EBI, you can search for a GO term to see its relationships and definition, as well as any available mappings to SWISS-PROT keywords, to the Enzyme Classification or Transport Classification databases, or to InterPro entries. Use documentation is available from the manual and the FAQ.
EP GO Browser	• The EP:GO browser is built into EBI's Expression Profiler, a set of tools for clustering, analysis and visualization of gene expression and other genomic data. With it, you can search for GO terms and identify gene associations for a node, with or without associated subnodes, for the organism of your choice.
GoFish	• The GoFish program, available as a Java applet, allows the user to construct arbitrary Boolean queries using GO attributes, and orders gene products according to the extent they satisfy such queries. GoFish also estimates, for each gene product, the probability that they satisfy the Boolean query. Developed by the Roth lab at Harvard.
GenNav	• GenNav is a GO browser developed at NLM. It searches GO terms and annotated gene products, and provides a graphical display of a term's position in the GO DAG.
GeneOntology@RZPD	• With the GeneOntology@RZPD tool at the Resource Center/Primary Database (RZPD) in Germany, you can search for GO identifiers associated with UniGene ClusterIds, Genes (Name/Symbol) and Clones provided by the RZPD. You can also search for UniGene Clusters, Genes and Clones annotated with a certain GO identifier or a combination of GO identifiers. So far, GO annotations for human and mouse genes/clones are linked.
ProToGO	• ProToGO, developed at the Hebrew University in Jerusalem, searches the GOA@EBI and Compugen annotation datasets. The output is a graphical view of the relevant sub-graph of GO, containing those GO terms assigned to the query proteins. Documentation is provided.
CGAP GO Browser	• With the GO browser at the The Cancer Genome Anatomy Project, you can browse through the GO vocabularies, and find human and mouse genes assigned to each term. The help documentation is at: http://cgap.nci.nih.gov/Genes/AllAboutGO.
DAG-Edit	
DAG-Edit	This Java application provides an interface to browse, query and edit GO or any other vocabulary that has a DAG data structure. The most current version of DAG-Edit can be downloaded from the publicly accessible source repository at SourceForge. Help documentation to use the program can also be downloaded from this site (.pdf or .html formats) or is available here: http://www.geneontology.org/doc/dagedit_userguide/dagedit.html
GO Database	
GO Database	API documentation, schema diagrams and full descriptions of all tables for the mySQL database developed and maintained by BDGP, http://www.godatabase.org/dev/database/

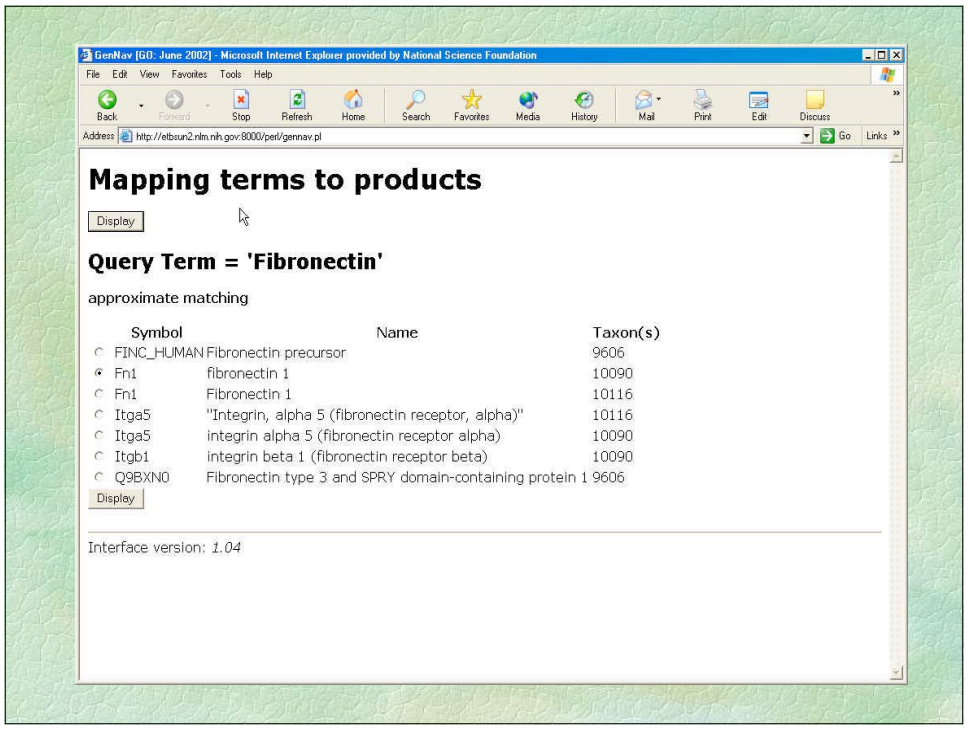
GO Database	
GO Database	API documentation, schema diagrams and full descriptions of all tables for the mySQL database developed and maintained by BDGP, http://www.godatabase.org/dev/database/

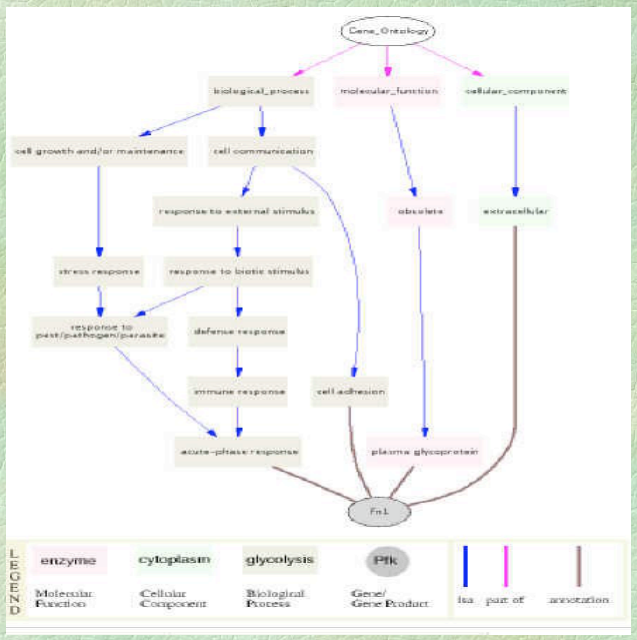
Other GO Tools	
GO Term Finder	• The GO Term Finder at SGD searches for significant shared GO terms, or parents of the GO terms, used to annotate budding yeast gene products.
GO Term Mapper	• The GO Term Mapper at SGD maps the specific, granular GO terms used to annotate a list of budding yeast gene products to corresponding GO Slim terms (i.e. more general parent GO terms; uses the SGD GO Slim set).
Manatee	• Manatee is a web-based gene evaluation and genome annotation tool developed at TIGR. Manatee can store and view annotation for prokaryotic and eukaryotic genomes. The Manatee interface allows biologists to quickly identify genes and make high quality functional assignments, such as GO classifications, using search data, paralogous families, and annotation suggestions generated from automated analysis.
PubSearch	• PubSearch is a web-based literature curation tool developed at TAIR and available via GMOD. It allows curators to search and annotate genes to keywords from articles. It has a simple, mySQL database backend and uses a set of Java Servlets and JSPs for querying, modifying, and adding gene, gene-annotation, and literature information. A demo is available.
SOURCE	• SOURCE, developed by the Stanford Microarray Database (SMD) team, compiles information from several publicly accessible databases, including UniGene, dbEST, Swiss-Prot, GeneMap99, RHdb, GeneCards and LocusLink. GO terms associated with LocusLink entries appear in SOURCE.
MAPPFinder	• MAPPFinder is an accessory program for GenMAPP. This program allows users to query any existing GenMAPP Expression Dataset Criterion against GO gene associations and GenMAPP MAPPs (microarray pathway profiles). The resulting analysis provides the user with results that can be viewed directly upon the Gene Ontology hierarchy and within GenMAPP, by selecting terms or MAPPs of interest.
FatiGO	• FatiGO is a web interface for clustering DNA microarray data and simple datamining using GO. datamining consists of the assignment of the most characteristic Gene Ontology term to a cluster. GO terms are related to Unigene Human and Mouse Cluster Ids and Saccharomyces Genome Database.
Onto-Express	• Onto-Express searches the public databases and returns tables that correlate expression profiles with the cytogenetic gene locations, the biochemical and molecular functions, the biological processes, cellular components and cellular roles of the translated proteins. (Registration required; free for academics.)
Genes2Diseases	• Genes2Diseases is a database of candidate genes for mapped inherited human diseases, developed by the Bork group at the European Molecular Biology Laboratory (EMBL). The database is generated using an analysis of relations between phenotypic features and chemical objects, and from chemical objects to protein function (Gene Ontology) terms, based on the whole MEDLINE and RefSeq databases. Can be used to view all GO terms associated with a particular genetically inherited disease.


Quick Search:
Submit
Site Map | Help | Full Search | Home
Community Info | Submit Data | BLAST | Primers | PatMatch | Gene/Seq Resources | Virtual Library | Contact SGD
RPE1/YJL121C
Help
Alternative single page format
RPE1 BASIC INFORMATION
Standard Name RPE1
Alias EPI1; POS18
Systematic Name YJL121C
Feature Type ORF
GO Annotations
RPE1 GO evidence and references
Molecular Function
ribulose-phosphate 3-epimerase activity
Biological Process
pentose-phosphate shunt
Cellular Component
cytosol
Description D-ribulose-5-Phosphate 3-epimerase
Gene Product D-ribulose-5-Phosphate 3-epimerase
Phenotype
RPE1 Phenotype details and references
Order mutant strains used in the systematic deletion project
Systematic deletion
viable
Exhibits growth defect on a non-fermentable (respiratory) carbon source.
Free text
Null mutants are viable but show no ribulose-5-phosphate epimerase activity, cannot grow on D-xylulose, and are sensitive to hydrogren peroxide
Sequence Coordinates ChrX: coordinates 190790 to 190074 [ORF Map]
Coordinates Chromosomal Coord
Exon 1 - 717 190790 - 190074
Note: this gene is encoded on the Crick strand.
External Links MIPS | YPD | SwissProt | Entrez Protein | Entrez Neighbors | PIR-JP | PIR-US | Entrez RefSeq | NiceZyme | Kyoto
Primary SGDID S0003657
RPE1 RESOURCES
Click on map for expanded view
Literature
Literature Guide View
Retrieve Sequences
Genomic DNA Retrieve
Sequence Analysis Tools
BLASTP Analyze
Protein Info & Structure
Protein Info View
Interactions
Interactions DB (GRID) View
Maps and Displays
Chr. Features Map View
Comparison Resources
Fungal alignments View
Functional Analysis
Alpha Factor Concentration View
ADDITIONAL INFORMATION for RPE1
Locus History | Global Gene Hunter | Function Junction | Expression Connection
Protein Info & Composition | Motifs | PDB Homologs | Gene/Sequence Resources
Return to SGD
Send a Message to the SGD Curators


BLAST Primers PatMatch Gene/Seq Resources Virtual Library Contact SGD
Gene Ontology: cytosol
Help
GO Tutorial
cytosol
chaperonin ATPase complex
cytosolic DNA-directed RNA polymerase (sensu Bacteria)
steroid hormone aporeceptor complex
hemoglobin complex
fatty-acyl-CoA synthase complex
26S proteasome 8(+36) genes
fatty-acid synthase complex 2 genes
cytosolic ribosome (sensu Bacteria)
IkappaB kinase complex
6-phosphofructokinase complex 2 genes
19S proteasome regulatory particle 20(+8) genes
20S core proteasome complex 15 genes
fatty acid elongase complex
cytosolic small ribosomal subunit (sensu Bacteria)
cytosolic large ribosomal subunit (sensu Bacteria)
19S proteasome regulatory particle, base subcomplex 4 genes
19S proteasome regulatory particle, lid subcomplex 4 genes
20S core proteasome, alpha-subunit complex
20S core proteasome, beta-subunit complex
proteasome activator complex


GenNav [GO: June 2002] - Microsoft Internet Explorer provided by National Science Foundation
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://etbsun2.nlm.nih.gov:8000/perl/gennav.pl
Go Links
GenNav
Term: Fibronectin
What: Gene/Gene product
Field: Automatic
String matching method: Automatic
Species restriction: none
Search GO Clear
GenNav will be temporarily unavailable on March 31, 2003
due to maintainance on the database server.
We apologize for any inconvenience.
Comments? Feedback? Questions?
Contact: Olivier Bodenreider
Copyright and Privacy Notice
Interface version: 1.04
Start
NSF WinStation
Sequence Alig...
SpengleMay1...
Inbox - Micros...
MGI 2.95 - Mo...
Saccharomyce...
GenNav [G...
11:49 AM

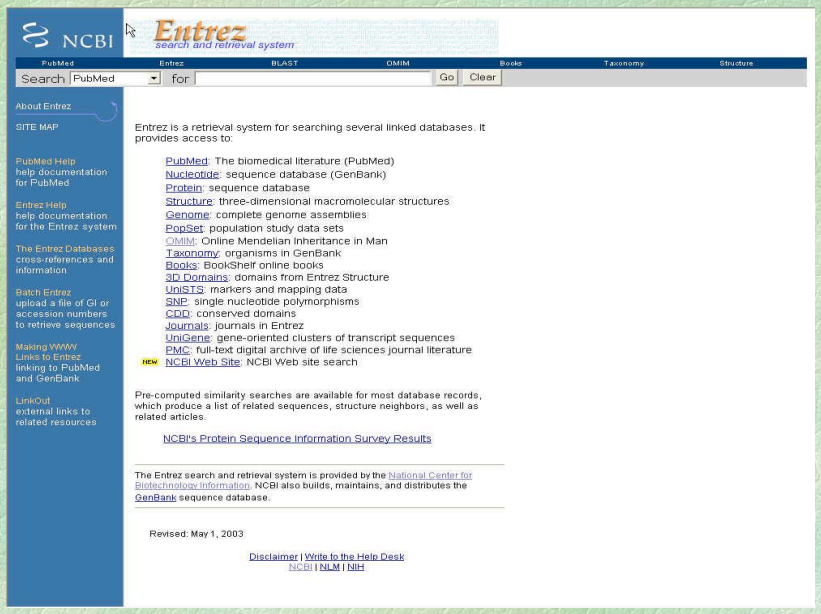

GenNav [GO: June 2002] - Microsoft Internet Explorer provided by National Science Foundation
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://etbsun2.nlm.nih.gov:8000/perl/gennav.pl
Go Links
Mapping terms to products
Display
Query Term = 'Fibronectin'
approximate matching
Symbol Name Taxon(s)
FINC_HUMAN Fibronectin precursor 9606
Fn1 fibronectin 1 10090
Fn1 Fibronectin 1 10116
Itga5 "Integrin, alpha 5 (fibronectin receptor, alpha)" 10116
Itga5 integrin alpha 5 (fibronectin receptor alpha) 10090
Itgb1 integrin beta 1 (fibronectin receptor beta) 10090
Q9BXN0 Fibronectin type 3 and SPRY domain-containing protein 1 9606
Display
Interface version: 1.04



Additional Information

Symbol	Fn1
Species	0
Full name	fibronectin 1
Synonyms	• Fn-1
Cross-references	DB TYPE ID OBJ_TYPE TAXON mgi acc MGI:95566 gene 10090
Molecular functions	• plasma glycoprotein [IEA]
Biological processes	• cell adhesion [IEA] • acute-phase response [IEA]
Cellular components	• extracellular [IEA]

New query



ENTREZ: Linked Databases

http://www.ncbi.nlm.nih.gov/Entrez/

- Concept of Neighbor-usually BLAST relationship
- Precomputed=Fast
- Related sequence, structure neighbors, related articles

- CUBBY

EST Pluses

- Rapid
- Inexpensive
- Applicable to gene discovery, regulation
- Gene sequence diversity

Worked to identify over half the human genes

EST negatives

- No defined protein product
- High error rate
- Not curated for annotation
- Low data quality for assembly

EST Data Quality Issues

- Regions of high and low quality reads in the same EST
- Kinds of error
  - Clone orientation
  - Chimeras
  - Missing reads
  - Compression and base calling

Stringent Clustering

- One-pass assembly
- Fewer, shorter consensus sequences
- Lower coverage of expressed gene data
- Lower inclusion rate of expressed gene forms
- Algorithms such as TIGR_ASSEMBLER

Loose Clustering

- One-pass assembly
- Large, “sloppy” clusters
- Greater coverage
- Possible inclusion of paralogous genes
- Lower fidelity
- Includes alternate expressed forms
- Unigene

Steps in EST Clustering

1. Mask for repeats and vector, leaving a minimum number of residues of ‘clean” data (100bp for Unigene)
2. Initial cluster based on sequence identity
3. Generation of consensus
4. Joining of clusters

STEPS IN CLUSERING-2

- Clone joining
  - Utilizes the physically shared clone id between 3' and 5' EST fragments from the same clone
    - Can be a source of error because it relies on the accuracy of the annotation and uniqueness of the clone ID, especially when data from disparate sources

What are we trying to do with EST Clustering?

- "Match" sequences-either they do or don't
- Look at near-identical matches

Tools for Matching

- Smith-Waterman, BLAST, FASTA built for searching-measure *quantitatively* the similarity between any two distances.
- But ESTs either match or not, so need only a near or perfect match.

Adjusting BLAST for ESTs

- Stringent match set for EST
  - E expectation value set to 0
  - G cost to open a gap increased
  - E cost to extend a gap increased
  - Q mismatch penalty increased
  - R match reward increased
  - W word size set for longer words

EST DATABASES:Quality issues

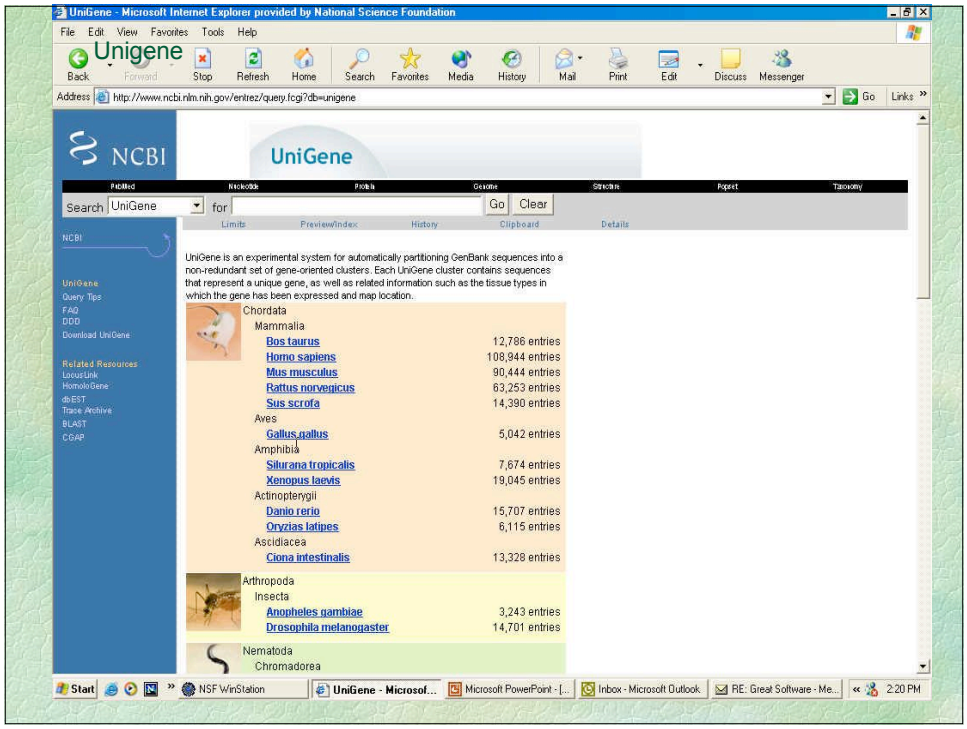
- SEQUENCE QUALITY
  - calculated error less than 1% (Phred-20) is the rule
  - frameshifts and stops common
  - Rules are usually observed by exception
  - There are lots of exceptions in the public data
  - Many 3' UTRs

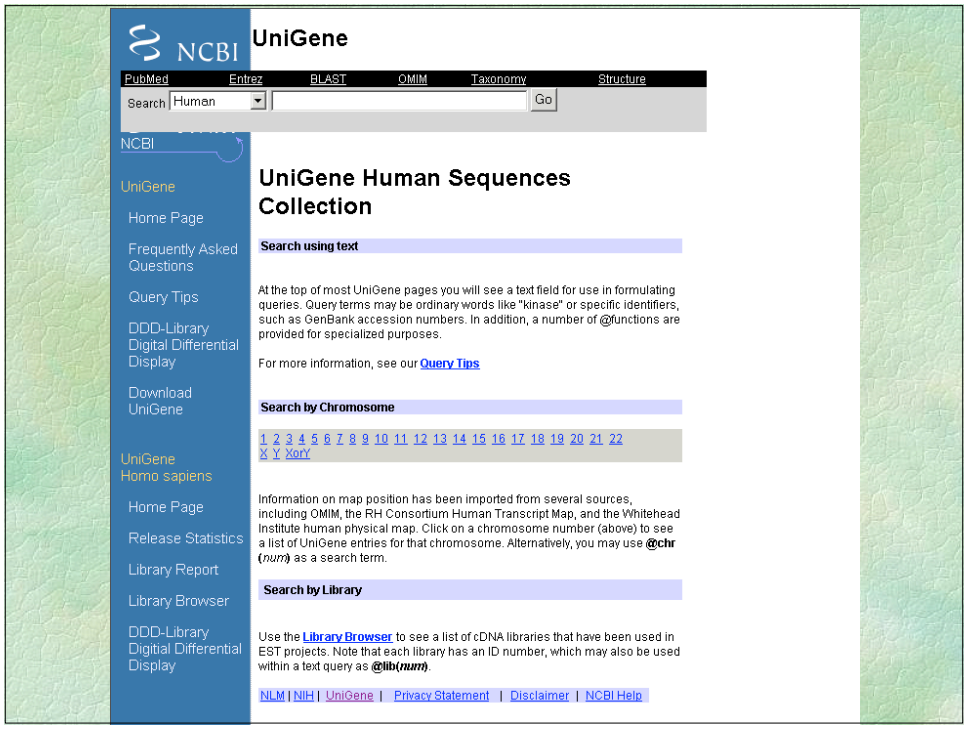
EST Databases: Quality #2

- CLONE QUALITY
  - Over-representation
  - Tissue specificity
  - Developmental stage specificity
  - Unprocessed mRNA clones
  - Chimeras
  - Contamination

EST Cluster Databases

- STACK-at SANBI http://sanbi.ac.za
- TIGR-animals, plants, other http://www.tigr.org/tdb/tgi.shtml
- Unigene-NCBI
  - Many species, plant, animal, …..
  - mRNAs
  - predicted mRNAs





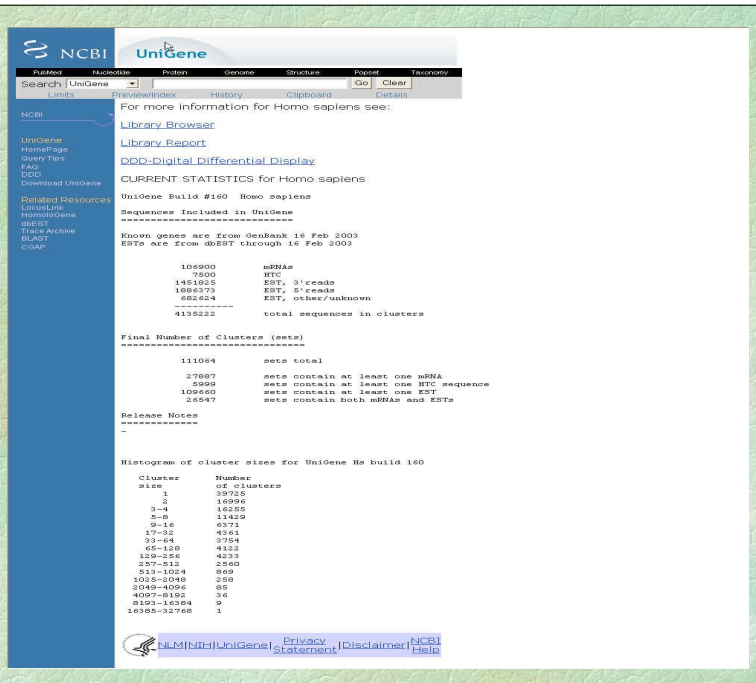
UNIGENE

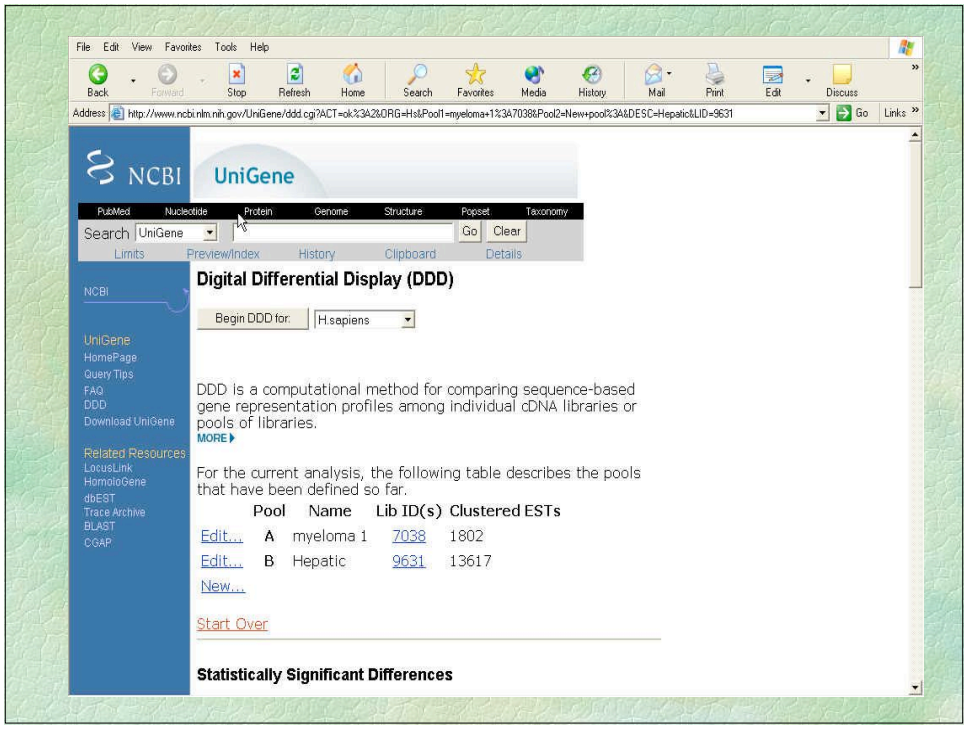
- A LIST OF LISTS
  - The cluster and known EST, mRNA pieces
  - Additional annotation-gene name, etc.
  - Distributed as a subset of dbest

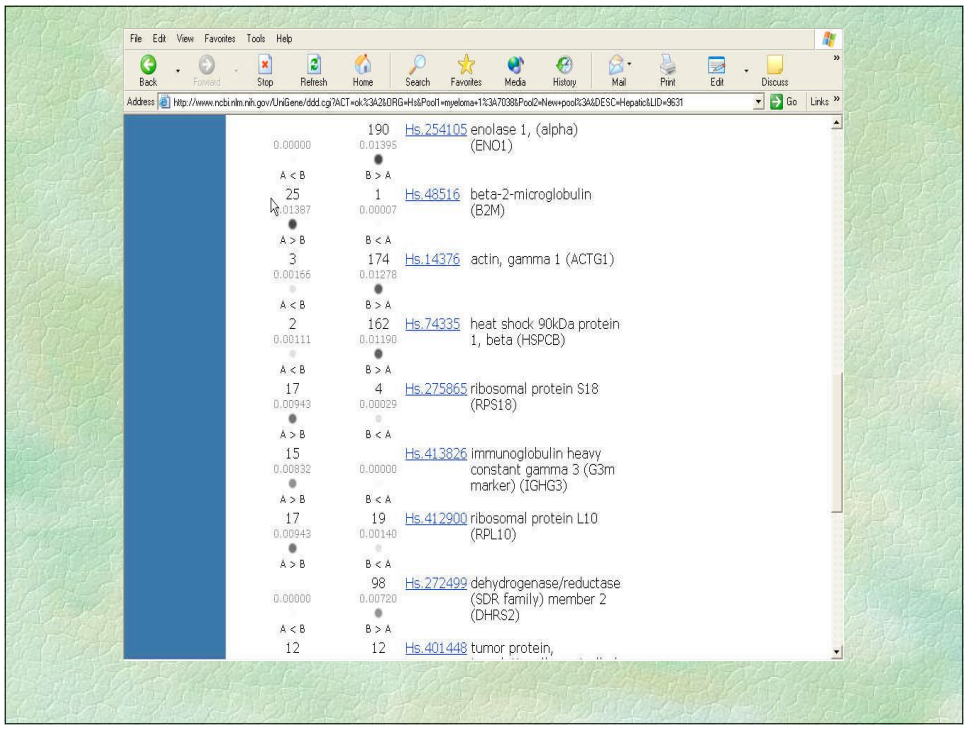
NOT included in the BLAST searchable DB at NCBI

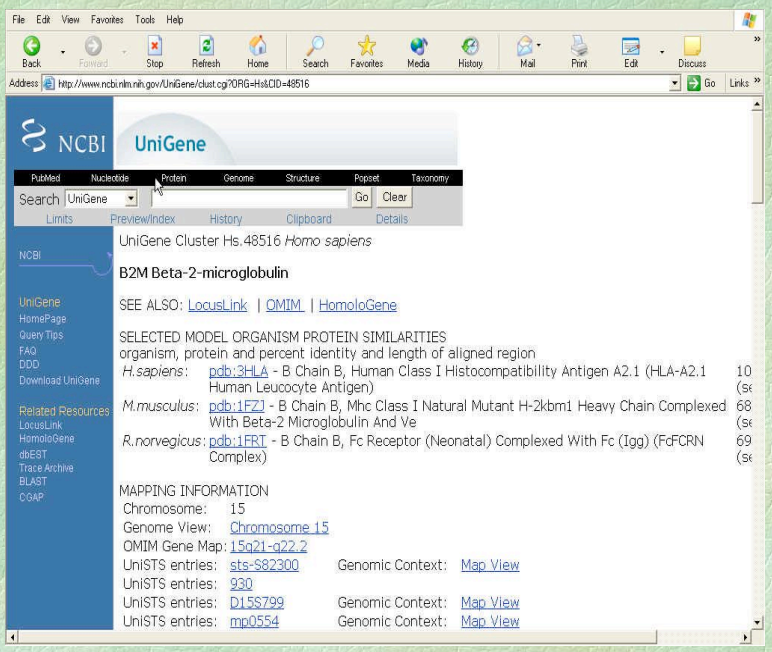
Caveats on Clusters

- Not stable
- Can go to complete cDNAs as available




File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ACT=ok%3A2&ORG=Hs&Pool1=myeloma+1%3A7038&Pool2=New+pool%3A&DESC=Hepatic&LID=9631
Go Links
NCBI UniGene
PubMed Nucleotide Protein Genome Structure Popset Taxonomy
Search UniGene Go Clear
Limits Preview/Index History Clipboard Details
NCBI
UniGene
HomePage
Query Tips
FAQ
DDD
Download UniGene
Related Resources
LocusLink
HomoloGene
dbEST
Trace Archive
BLAST
CGAP
Digital Differential Display (DDD)
Begin DDD for: H.sapiens
DDD is a computational method for comparing sequence-based gene representation profiles among individual cDNA libraries or pools of libraries.
MORE
For the current analysis, the following table describes the pools that have been defined so far.
Pool Name Lib ID(s) Clustered ESTs
Edit... A myeloma 1 7038 1802
Edit... B Hepatic 9631 13617
New...
Start Over
Statistically Significant Differences

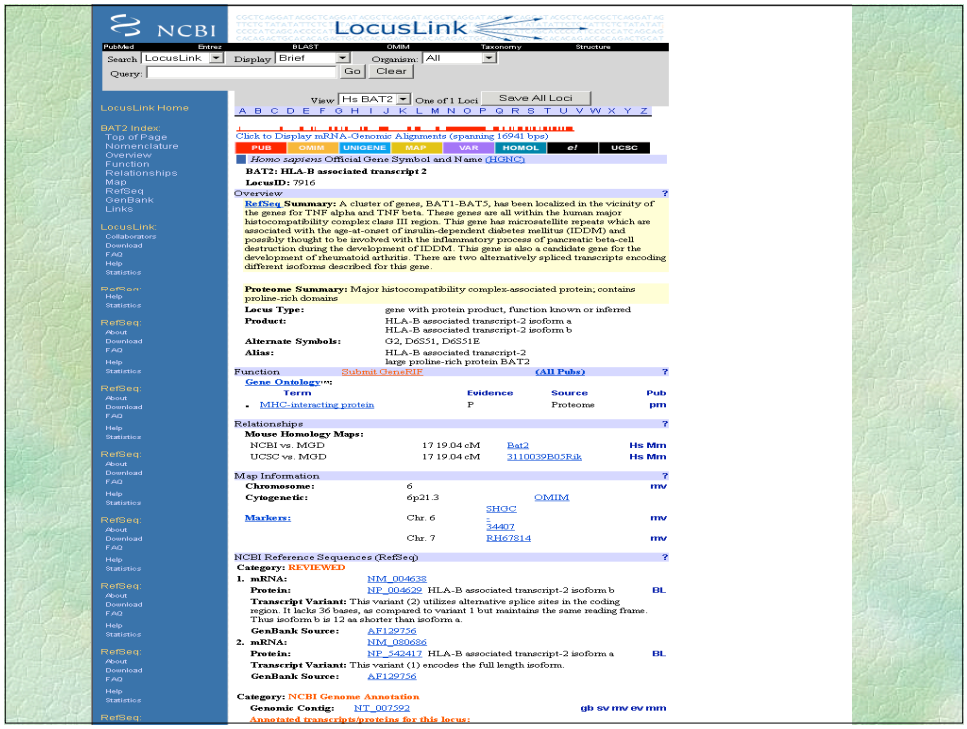

File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ACT=ok%3A2&ORG=Hs&Pool1=myeloma+1%3A7038&Pool2=New+pool%3A&DESC=Hepatic&LID=9631
Go Links
0.00000 A < B 190 0.01395 B > A Hs.254105 enolase 1, (alpha) (ENO1)
0.01387 A > B 25 1 0.00007 B < A Hs.48516 beta-2-microglobulin (B2M)
0.00166 A < B 3 174 0.01278 B > A Hs.14376 actin, gamma 1 (ACTG1)
0.00111 A < B 2 162 0.01190 B > A Hs.74335 heat shock 90kDa protein 1, beta (HSPCB)
0.00943 A > B 17 4 0.00029 B < A Hs.275865 ribosomal protein S18 (RPS18)
0.00832 A > B 15 0.00000 B < A Hs.413826 immunoglobulin heavy constant gamma 3 (G3m marker) (IGHG3)
0.00943 A > B 17 19 0.00140 B < A Hs.412900 ribosomal protein L10 (RPL10)
0.00000 A < B 98 0.00720 B > A Hs.272499 dehydrogenase/reductase (SDR family) member 2 (DHRS2)
12 12 Hs.401448 tumor protein,

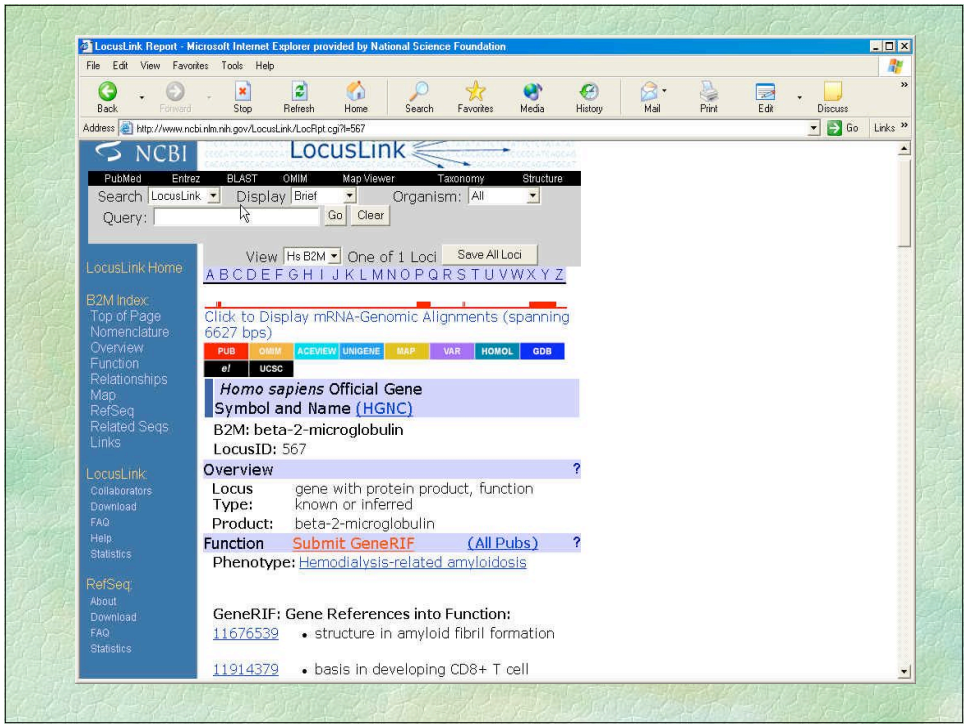


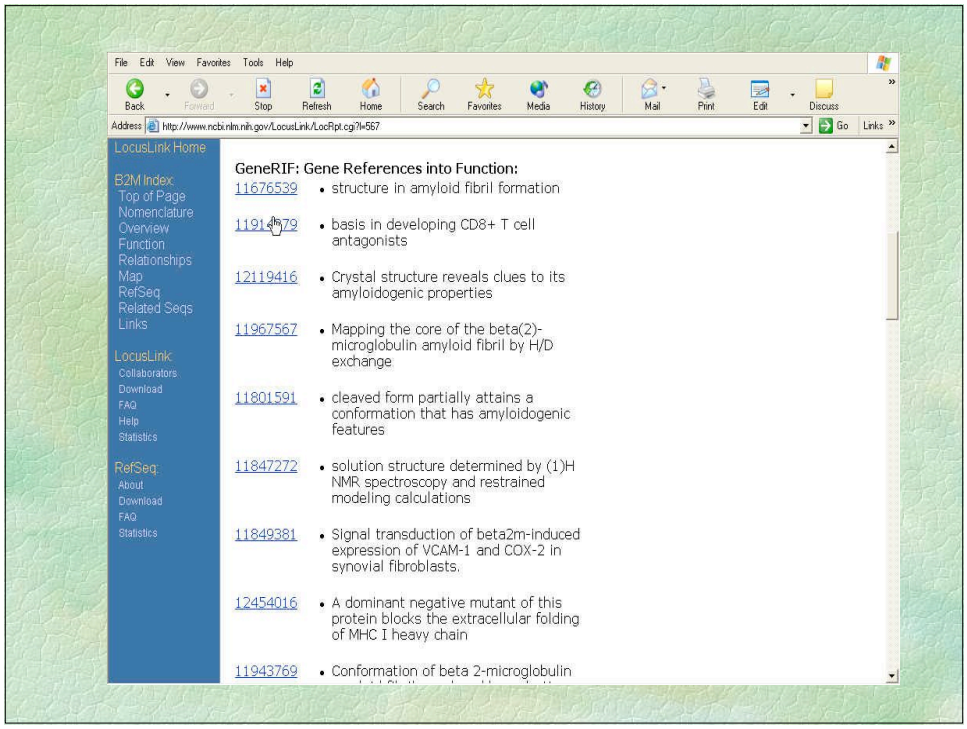
LOCUSLINK

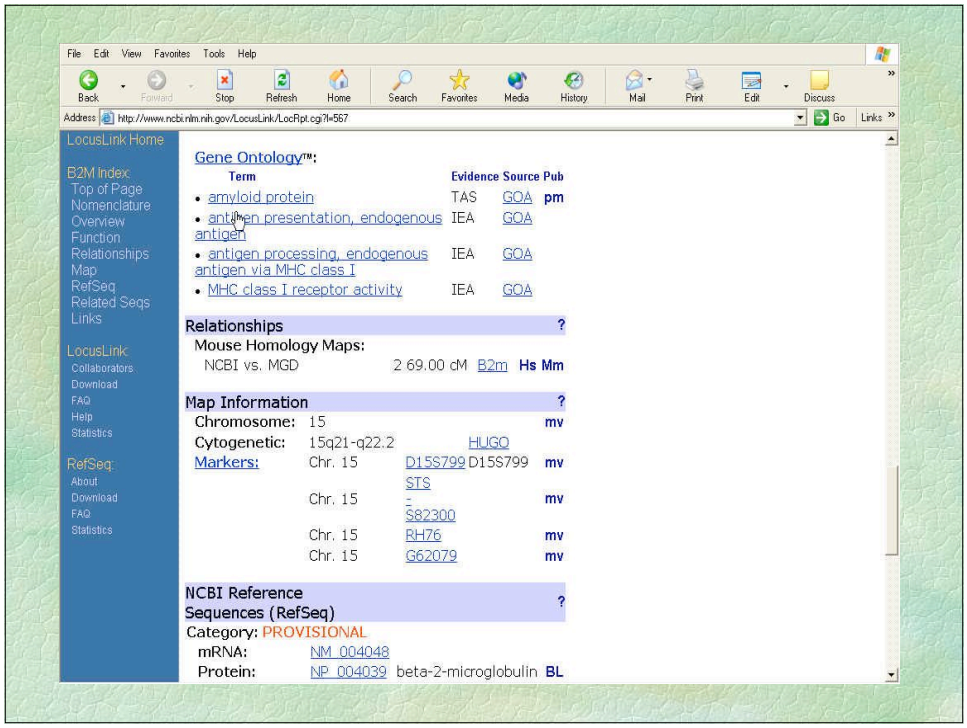
(http://www.ncbi.nlm.nih.gov/LocusLink)

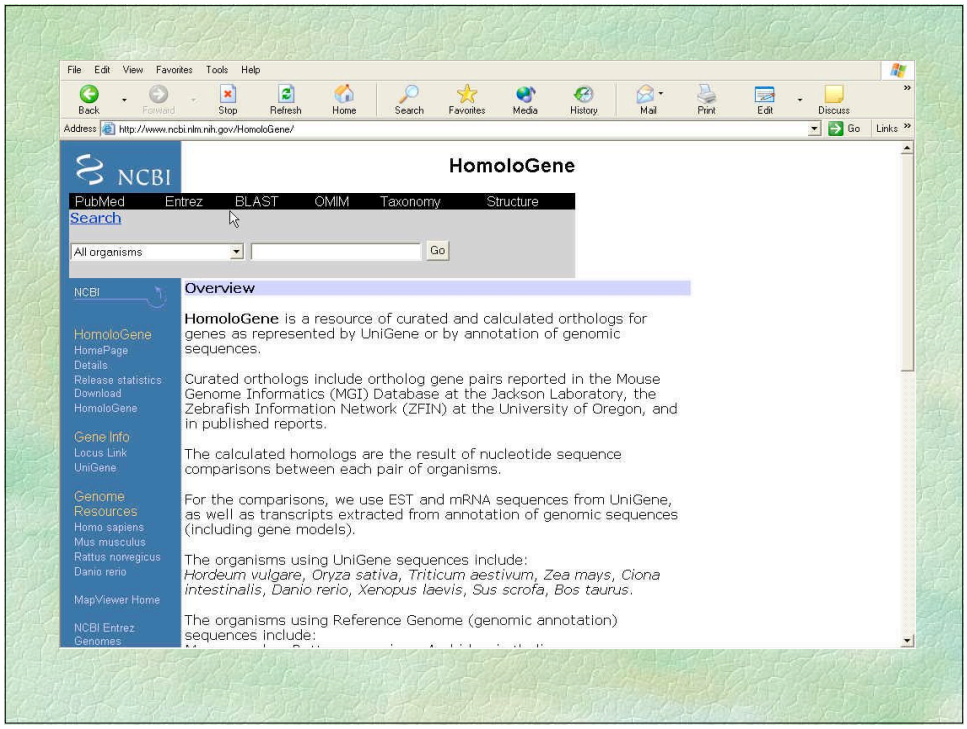
- A useful, searchable compendium of loci across *Caenorhabditis elegans*, cow, fruit fly, human, human immunodeficiency virus type 1, mouse, rat, and zebrafish.
- Linked for PubMed, OMIM, RefSeq, Homologene data, Unigene, and Variation Data

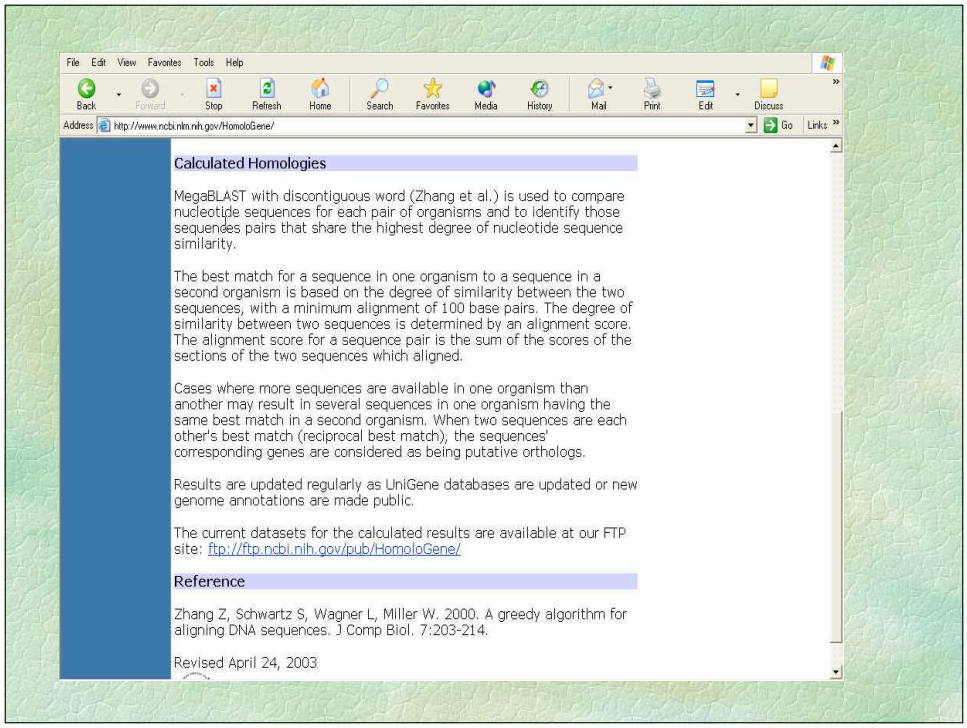

NCBI
LocusLink
Search LocusLink
Display Brief
Organism: All
Query:
Go
Clear
View Hs BAT2
One of 1 Loci
Save All Loci
LocusLink Home
BAT2 Index:
Top of Page
Nomenclature
Overview
Function
Relationships
Map
RefSeq
GenBank
Links
LocusLink:
Collaborators
Download
FAQ
Help
Statistics
RefSeq:
About
Download
FAQ
Help
Statistics
Click to Display mRNA-Genomic Alignments (spanning 16941 bps)
PUB
OMIM
UNIGENE
MAP
VAR
HOMOL
e!
UCSC
Homo sapiens Official Gene Symbol and Name (HGNC)
BAT2: HLA-B associated transcript 2
LocusID: 7916
Overview
RefSeq Summary: A cluster of genes, BAT1-BAT5, has been localized in the vicinity of the genes for TNF alpha and TNF beta. These genes are all within the human major histocompatibility complex class III region. This gene has microsatellite repeats which are associated with the age-at-onset of insulin-dependent diabetes mellitus (IDDM) and possibly thought to be involved with the inflammatory process of pancreatic beta-cell destruction during the development of IDDM. This gene is also a candidate gene for the development of rheumatoid arthritis. There are two alternatively spliced transcripts encoding different isoforms described for this gene.
Proteome Summary: Major histocompatibility complex-associated protein; contains proline-rich domains
Locus Type: gene with protein product, function known or inferred
Product: HLA-B associated transcript-2 isoform a
HLA-B associated transcript-2 isoform b
Alternate Symbols: G2, D6S51, D6S51E
Alias: HLA-B associated transcript-2
large proline-rich protein BAT2
Function
Submit GeneRIF
(All Pubs)
Gene Ontology:
Term
Evidence
Source
Pub
MHC-interacting protein
F
Proteome
pm
Relationships
Mouse Homology Maps:
NCBI vs. MGD
17 19.04 cM
Bat2
Hs Mm
UCSC vs. MGD
17 19.04 cM
3110039B05Rik
Hs Mm
Map Information
Chromosome: 6
Cytogenetic: 6p21.3
OMIM
Markers:
Chr. 6
SHGC-34407
mv
Chr. 7
RH67814
mv
NCBI Reference Sequences (RefSeq)
Category: REVIEWED
1. mRNA: NM_004638
Protein: NP_004629 HLA-B associated transcript-2 isoform b
Transcript Variant: This variant (2) utilizes alternative splice sites in the coding region. It lacks 36 bases, as compared to variant 1 but maintains the same reading frame. Thus isoform b is 12 aa shorter than isoform a.
GenBank Source: AF129756
2. mRNA: NM_0306936
Protein: NP_542417 HLA-B associated transcript-2 isoform a
Transcript Variant: This variant (1) encodes the full length isoform.
GenBank Source: AF129756
Category: NCBI Genome Annotation
Genomic Contig: NT_007592
gb sv mv ev mm
Annotated transcripts/proteins for this locus:

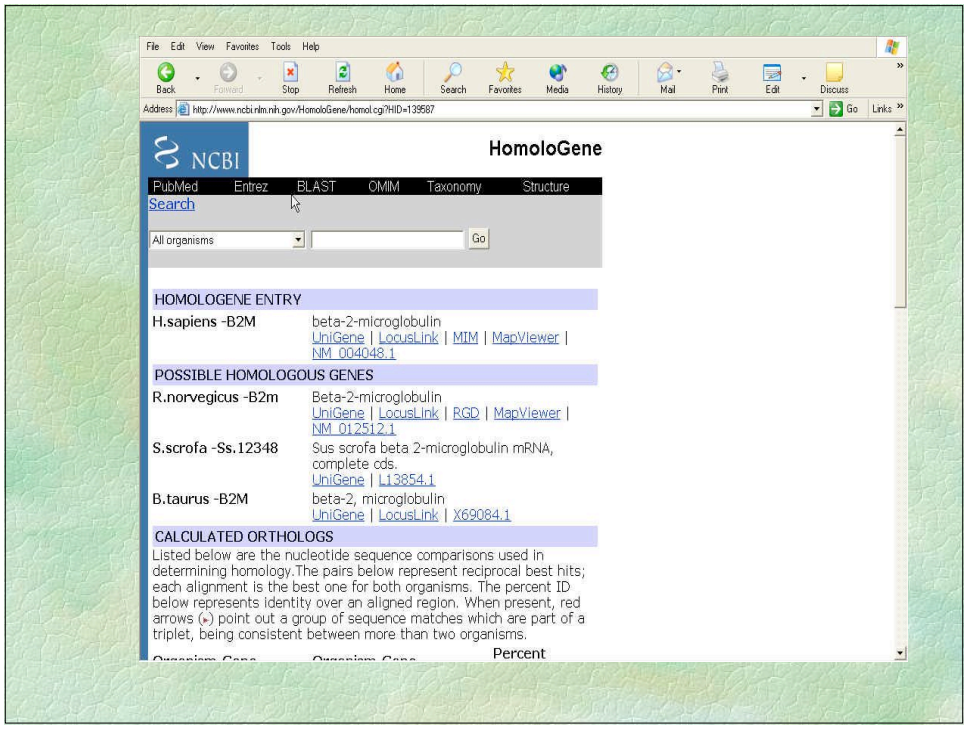

LocusLink Report - Microsoft Internet Explorer provided by National Science Foundation
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=567
Go Links
NCBI
LocusLink
PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure
Search LocusLink
Display Brief
Organism: All
Query:
Go
Clear
View Hs B2M
One of 1 Loci
Save All Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
LocusLink Home
B2M Index:
Top of Page
Nomenclature
Overview
Function
Relationships
Map
RefSeq
Related Seqs
Links
LocusLink:
Collaborators
Download
FAQ
Help
Statistics
RefSeq:
About
Download
FAQ
Statistics
Click to Display mRNA-Genomic Alignments (spanning 6627 bps)
PUB
OMIM
ACEVIEW
UNIGENE
MAP
VAR
HOMOL
GDB
e!
UCSC
Homo sapiens Official Gene Symbol and Name (HGNC)
B2M: beta-2-microglobulin
LocusID: 567
Overview
Locus Type: gene with protein product, function known or inferred
Product: beta-2-microglobulin
Function
Submit GeneRIF
(All Pubs)
Phenotype: Hemodialysis-related amyloidosis
GeneRIF: Gene References into Function:
11676539 • structure in amyloid fibril formation
11914379 • basis in developing CD8+ T cell

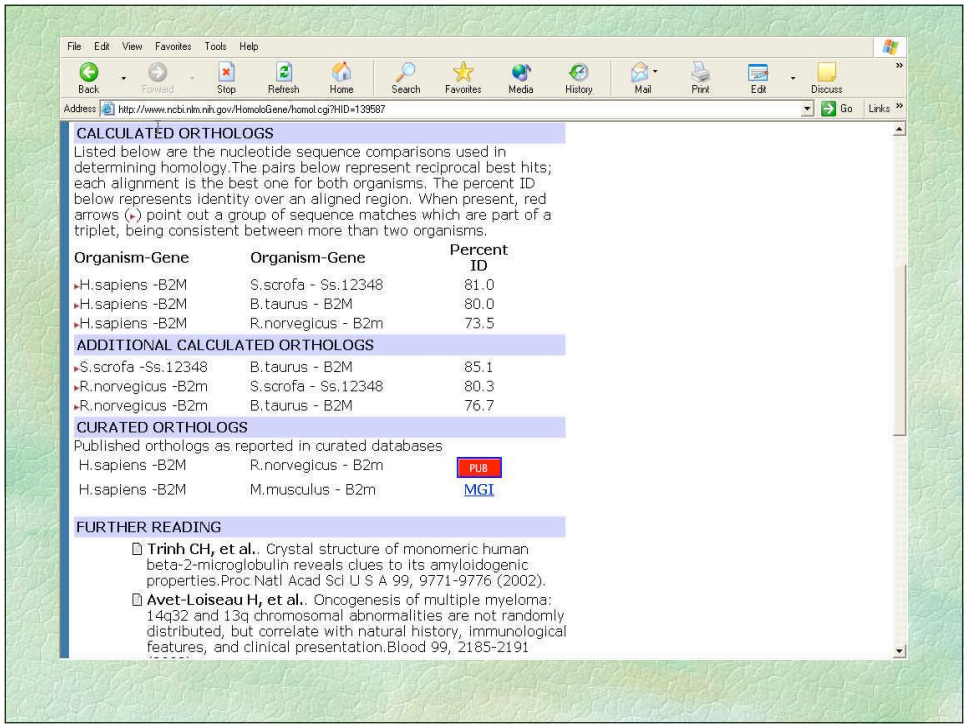

File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=567
Go Links
LocusLink Home
B2M Index:
Top of Page
Nomenclature
Overview
Function
Relationships
Map
RefSeq
Related Seqs
Links
LocusLink:
Collaborators
Download
FAQ
Help
Statistics
RefSeq:
About
Download
FAQ
Statistics
GeneRIF: Gene References into Function:
11676539 • structure in amyloid fibril formation
11914879 • basis in developing CD8+ T cell antagonists
12119416 • Crystal structure reveals clues to its amyloidogenic properties
11967567 • Mapping the core of the beta(2)-microglobulin amyloid fibril by H/D exchange
11801591 • cleaved form partially attains a conformation that has amyloidogenic features
11847272 • solution structure determined by (1)H NMR spectroscopy and restrained modeling calculations
11849381 • Signal transduction of beta2m-induced expression of VCAM-1 and COX-2 in synovial fibroblasts.
12454016 • A dominant negative mutant of this protein blocks the extracellular folding of MHC I heavy chain
11943769 • Conformation of beta 2-microglobulin

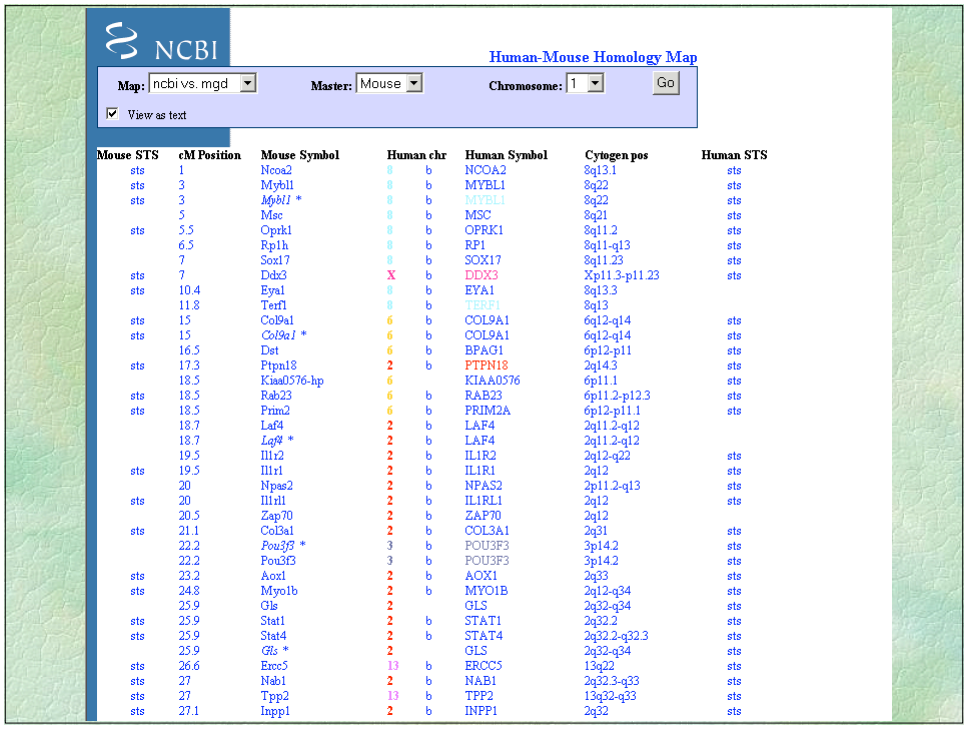

File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=567
Go Links
LocusLink Home
B2M Index:
Top of Page
Nomenclature
Overview
Function
Relationships
Map
RefSeq
Related Seqs
Links
LocusLink:
Collaborators
Download
FAQ
Help
Statistics
RefSeq:
About
Download
FAQ
Statistics
Gene Ontology™:
Term Evidence Source Pub
• amyloid protein TAS GOA pm
• antigen presentation, endogenous antigen IEA GOA
• antigen processing, endogenous antigen via MHC class I IEA GOA
• MHC class I receptor activity IEA GOA
Relationships ?
Mouse Homology Maps:
NCBI vs. MGD 2 69.00 cM B2m Hs Mm
Map Information ?
Chromosome: 15 mv
Cytogenetic: 15q21-q22.2 HUGO
Markers: Chr. 15 D15S799 D15S799 mv
Chr. 15 STS - S82300 mv
Chr. 15 RH76 mv
Chr. 15 G62079 mv
NCBI Reference Sequences (RefSeq) ?
Category: PROVISIONAL
mRNA: NM_004048
Protein: NP_004039 beta-2-microglobulin BL


File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/HomoloGene/
Go Links
NCBI
HomoloGene
PubMed Entrez BLAST OMIM Taxonomy Structure
Search
All organisms
Go
NCBI
HomoloGene
HomePage
Details
Release statistics
Download
HomoloGene
Gene Info
Locus Link
UniGene
Genome Resources
Homo sapiens
Mus musculus
Rattus norvegicus
Danio rerio
MapViewer Home
NCBI Entrez Genomes
Overview
HomoloGene is a resource of curated and calculated orthologs for genes as represented by UniGene or by annotation of genomic sequences.
Curated orthologs include ortholog gene pairs reported in the Mouse Genome Informatics (MGI) Database at the Jackson Laboratory, the Zebrafish Information Network (ZFIN) at the University of Oregon, and in published reports.
The calculated homologs are the result of nucleotide sequence comparisons between each pair of organisms.
For the comparisons, we use EST and mRNA sequences from UniGene, as well as transcripts extracted from annotation of genomic sequences (including gene models).
The organisms using UniGene sequences include:
Hordeum vulgare, Oryza sativa, Triticum aestivum, Zea mays, Ciona intestinalis, Danio rerio, Xenopus laevis, Sus scrofa, Bos taurus.
The organisms using Reference Genome (genomic annotation) sequences include:


File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/HomoloGene/
Go Links
Calculated Homologies
MegaBLAST with discontiguous word (Zhang et al.) is used to compare nucleotide sequences for each pair of organisms and to identify those sequences pairs that share the highest degree of nucleotide sequence similarity.
The best match for a sequence in one organism to a sequence in a second organism is based on the degree of similarity between the two sequences, with a minimum alignment of 100 base pairs. The degree of similarity between two sequences is determined by an alignment score. The alignment score for a sequence pair is the sum of the scores of the sections of the two sequences which aligned.
Cases where more sequences are available in one organism than another may result in several sequences in one organism having the same best match in a second organism. When two sequences are each other's best match (reciprocal best match), the sequences' corresponding genes are considered as being putative orthologs.
Results are updated regularly as UniGene databases are updated or new genome annotations are made public.
The current datasets for the calculated results are available at our FTP site: ftp://ftp.ncbi.nih.gov/pub/HomoloGene/
Reference
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comp Biol. 7:203-214.
Revised April 24, 2003


File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/HomoloGene/homol.cgi?HID=139587
Go Links
NCBI
HomoloGene
PubMed Entrez BLAST OMIM Taxonomy Structure
Search
All organisms
Go
HOMOLOGENE ENTRY
H.sapiens -B2M
beta-2-microglobulin
UniGene | LocusLink | MIM | MapViewer | NM_004048.1
POSSIBLE HOMOLOGOUS GENES
R.norvegicus -B2m
Beta-2-microglobulin
UniGene | LocusLink | RGD | MapViewer | NM_012512.1
S.scrofa -Ss.12348
Sus scrofa beta 2-microglobulin mRNA, complete cds.
UniGene | L13854.1
B.taurus -B2M
beta-2, microglobulin
UniGene | LocusLink | X69084.1
CALCULATED ORTHOLOGS
Listed below are the nucleotide sequence comparisons used in determining homology.The pairs below represent reciprocal best hits; each alignment is the best one for both organisms. The percent ID below represents identity over an aligned region. When present, red arrows (▸) point out a group of sequence matches which are part of a triplet, being consistent between more than two organisms.
Percent


File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/HomoloGene/homol.cgi?HID=139587
Go Links
CALCULATED ORTHOLOGS
Listed below are the nucleotide sequence comparisons used in determining homology.The pairs below represent reciprocal best hits; each alignment is the best one for both organisms. The percent ID below represents identity over an aligned region. When present, red arrows (▸) point out a group of sequence matches which are part of a triplet, being consistent between more than two organisms.
Organism-Gene Organism-Gene Percent ID
▸H.sapiens -B2M S.scrofa - Ss.12348 81.0
▸H.sapiens -B2M B.taurus - B2M 80.0
▸H.sapiens -B2M R.norvegicus - B2m 73.5
ADDITIONAL CALCULATED ORTHOLOGS
▸S.scrofa -Ss.12348 B.taurus - B2M 85.1
▸R.norvegicus -B2m S.scrofa - Ss.12348 80.3
▸R.norvegicus -B2m B.taurus - B2M 76.7
CURATED ORTHOLOGS
Published orthologs as reported in curated databases
H.sapiens -B2M R.norvegicus - B2m PUB
H.sapiens -B2M M.musculus - B2m MGI
FURTHER READING
Trinh CH, et al.. Crystal structure of monomeric human beta-2-microglobulin reveals clues to its amyloidogenic properties.Proc Natl Acad Sci U S A 99, 9771-9776 (2002).
Avet-Loiseau H, et al.. Oncogenesis of multiple myeloma: 14q32 and 13q chromosomal abnormalities are not randomly distributed, but correlate with natural history, immunological features, and clinical presentation.Blood 99, 2185-2191



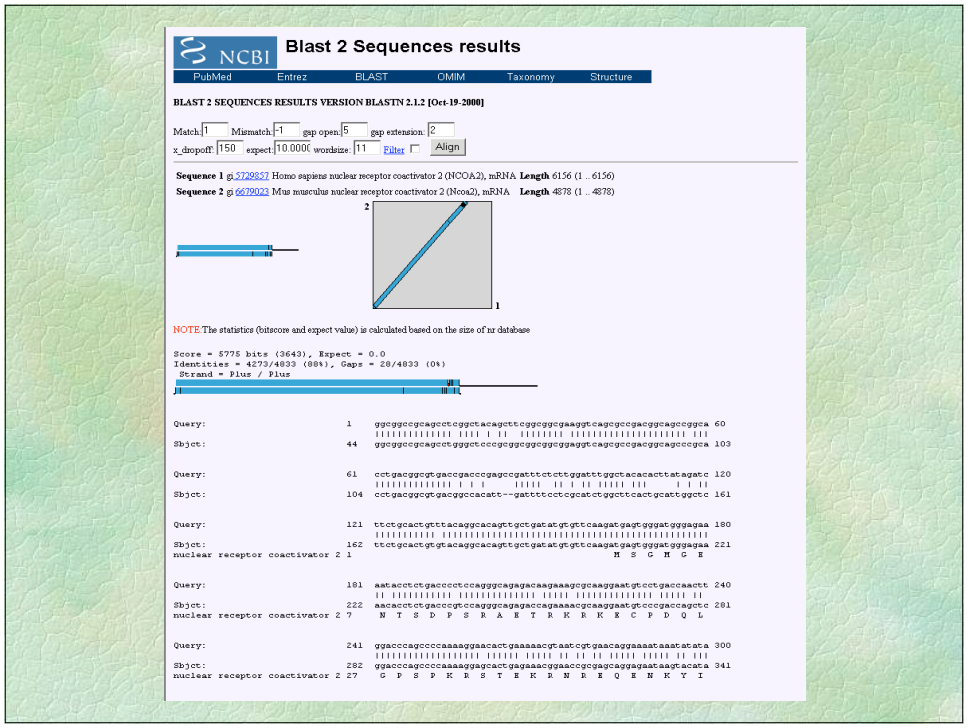
NCBI

Human-Mouse Homology Map

Map: ncbi vs. mgd Master: Mouse Chromosome: 1 Go

View as text

Mouse STS	cM Position	Mouse Symbol	Human chr		Human Symbol	Cytogen pos	Human STS
sts	1	Ncoa2	8	b	NCOA2	8q13.1	sts
sts	3	Mybl1	8	b	MYBL1	8q22	sts
sts	3	*Mybl1* *	8	b	MYBL1	8q22	sts
	5	Msc	8	b	MSC	8q21	sts
sts	5.5	Oprk1	8	b	OPRK1	8q11.2	sts
	6.5	Rp1h	8	b	RP1	8q11-q13	sts
	7	Sox17	8	b	SOX17	8q11.23	sts
sts	7	Ddx3	X	b	DDX3	Xp11.3-p11.23	sts
sts	10.4	Eya1	8	b	EYA1	8q13.3	
	11.8	Terf1	8	b	TERF1	8q13	
sts	15	Col9a1	6	b	COL9A1	6q12-q14	sts
sts	15	*Col9a1* *	6	b	COL9A1	6q12-q14	sts
	16.5	Dst	6	b	BPAG1	6p12-p11	sts
sts	17.3	Ptpn18	2	b	PTPN18	2q14.3	sts
	18.5	Kiaa0576-hp	6		KIAA0576	6p11.1	sts
sts	18.5	Rab23	6	b	RAB23	6p11.2-p12.3	sts
sts	18.5	Prim2	6	b	PRIM2A	6p12-p11.1	sts
	18.7	Laf4	2	b	LAF4	2q11.2-q12	
	18.7	*Laf4* *	2	b	LAF4	2q11.2-q12	
	19.5	Il1r2	2	b	IL1R2	2q12-q22	sts
sts	19.5	Il1r1	2	b	IL1R1	2q12	sts
	20	Npas2	2	b	NPAS2	2p11.2-q13	sts
sts	20	Il1rl1	2	b	IL1RL1	2q12	sts
	20.5	Zap70	2	b	ZAP70	2q12	
sts	21.1	Col3a1	2	b	COL3A1	2q31	sts
	22.2	*Pou3f3* *	3	b	POU3F3	3p14.2	sts
	22.2	Pou3f3	3	b	POU3F3	3p14.2	sts
sts	23.2	Aox1	2	b	AOX1	2q33	sts
sts	24.8	Myo1b	2	b	MYO1B	2q12-q34	sts
	25.9	Gls	2		GLS	2q32-q34	sts
sts	25.9	Stat1	2	b	STAT1	2q32.2	sts
sts	25.9	Stat4	2	b	STAT4	2q32.2-q32.3	sts
	25.9	*Gls* *	2		GLS	2q32-q34	sts
sts	26.6	Ercc5	13	b	ERCC5	13q22	sts
sts	27	Nab1	2	b	NAB1	2q32.3-q33	sts
sts	27	Tpp2	13	b	TPP2	13q32-q33	sts
sts	27.1	Inpp1	2	b	INPP1	2q32	sts



NCBI Blast 2 Sequences results

PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.1.2 [Oct-19-2000]

Match: 1 Mismatch: -1 gap open: 5 gap extension: 2

x_dropoff: 150 expect: 10.0000 wordsize: 11 Filter Align

Sequence 1 gi|5729857 Homo sapiens nuclear receptor coactivator 2 (NCOA2), mRNA **Length** 6156 (1 .. 6156)

Sequence 2 gi|6679023 Mus musculus nuclear receptor coactivator 2 (Ncoa2), mRNA **Length** 4878 (1 .. 4878)

NOTE The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score = 5775 bits (3643), Expect = 0.0
Identities = 4273/4833 (88%), Gaps = 28/4833 (0%)
 Strand = Plus / Plus

Query:                          1    ggcggccgcagcctcggctacagcttcggcggcgaaggtcagcgccgacggcagccggca 60
                                     ||||||||||||||| |||| | ||  |||||||| |||||||||||||||||||||| |||
Sbjct:                          44   ggcggccgcagcctgggctcccgcgggcggcggcggaggtcagcgccgacggcagcccgca 103

Query:                          61   cctgacggcgtgaccgacccgagccgatttctcttggatttggctacacacttatagatc 120
                                     |||||||||||||| | | |      |||||  || | || ||||| |||     | | ||
Sbjct:                          104  cctgacggcgtgacggccacatt--gattttcctcgcatctggcttcactgcattggctc 161

Query:                          121  ttctgcactgtttacaggcacagttgctgatatgtgttcaagatgagtgggatgggagaa 180
                                     |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                          162  ttctgcactgtgtacaggcacagttgctgatatgtgttcaagatgagtgggatgggagaa 221
nuclear receptor coactivator 2 1                                      M  S  G  M  G  E

Query:                          181  aatacctcttgaccctccagggcagagacaagaaagcgcaaggaatgtcctgaccaactt 240
                                     || |||||||||||| |||||||||||||||| ||||| ||||||||||||||||| ||||| ||
Sbjct:                          222  aacacctctgaccgctccagggcagagaccagaaaacgcaaggaatgtccgaccagctc 281
nuclear receptor coactivator 2 7    N  T  S  D  P  S  R  A  E  T  R  K  R  K  E  C  P  D  Q  L

Query:                          241  ggacccagcccaaaaggaacactgaaaaacgtaatcgtgaacaggaaaataaatatata 300
                                     |||||||||||||||||||| |||||| ||||| || || || ||||| ||||| || |||
Sbjct:                          282  ggacccagccccaaaaggagcactgagaaacggaaccgcgagcaggagaataagtacata 341
nuclear receptor coactivator 2 27   G  P  S  P  K  R  S  T  E  K  R  N  R  E  Q  E  N  K  Y  I
```

Resources for Genomic Comparison

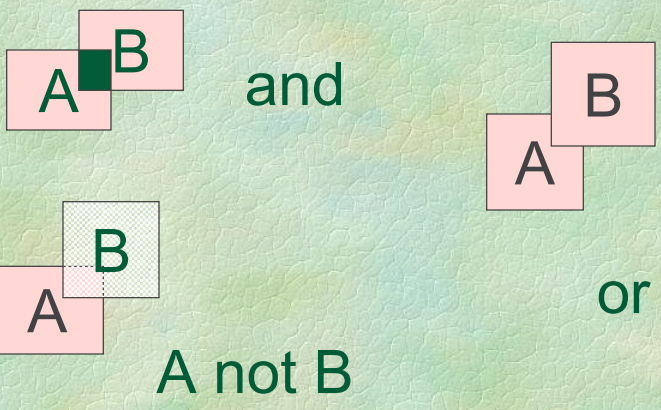
- GLASS-http://plover.lcs.mit.edu
- PipMaker: http://bio.cse.psu.edu
- Rosetta: http:// plover.lcs.mit.edu(genes)
- SGP: htttp://soft.ice.mpg.de/sgp-1
- VISTA: http://www-gsd.lbl.gov/VISTA
- WABA: http://www.cse.ucsc.edu/~kent/xenoAli/index.html

EFFICIENT TEXT SEARCHING

- Use Wild Cards: #,$,?,*
- Use Boolean Operators
  - Not
  - And
  - Or
  - Nor

Boolean Operators

- *AND* A and B BOTH
- *OR* A or B EITHER
- *NOT* B not A Have B, do not have A
- *NOR* A nor B A but not B OR B but not A



WILD CARDS

- Match one character-NCBI uses #
- Match zero or one character NCBI uses $, others ?
- Match zero or more characters-usually *

MEDICAL SUBJECT HEADINGS

- CONTROLLED Vocabulary
- Indexing of articles, books, etc.
- Current version has over 300,000 terms
- Can download list and make your own assortment

MeSH Advantages

- Assigned to the the entire document, not just title and abstract
- Major topic (*)
- Subheadings if available
- MeSH topics are exploded to include all the terms included in the meaning.

Try it; you may like it.

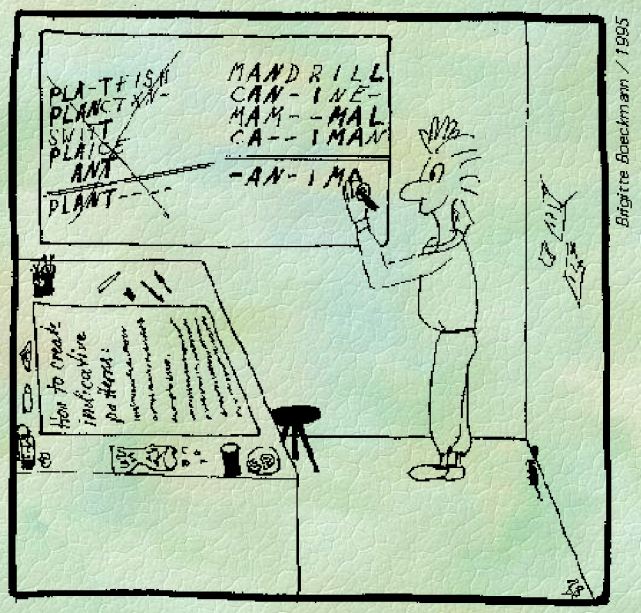
Other Resources

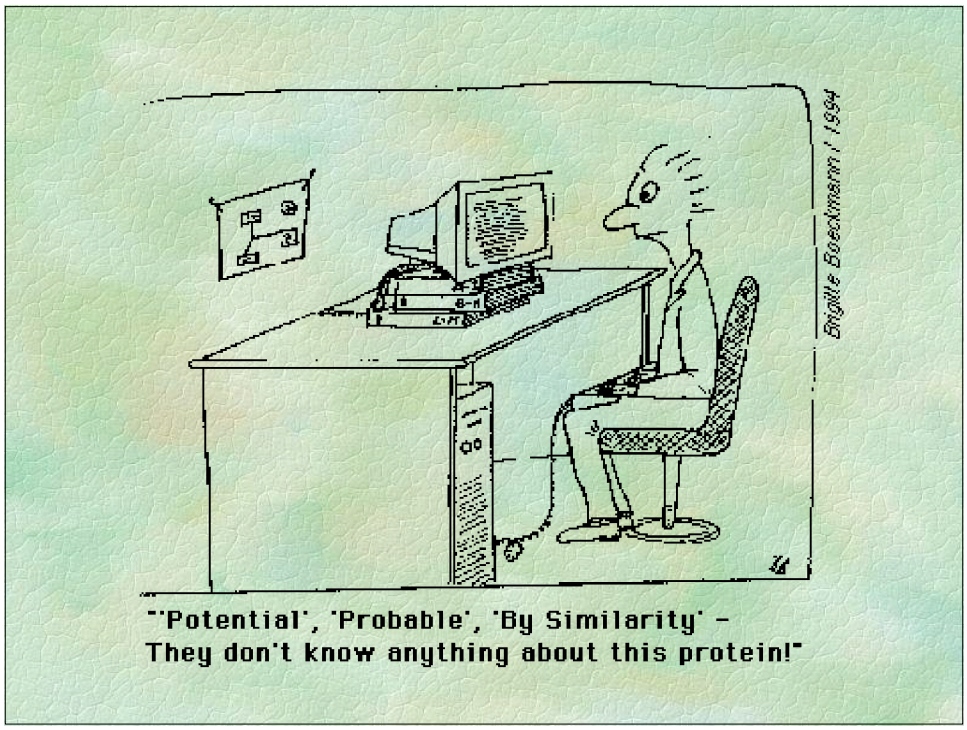
- NCBI Education Page http://www.ncbi.nlm.nih.gov/Education/index.html
- BCM Gene Finder http://searchlauncher.bcm.tmc.edu/docs/sl_links.html
- EBI-SwissProt, TrEMBL, PIR, SRS, Tools http://www.ebi.ac.uk
- ExPASy-SwissProt, TrEMBL http://www.expasy.ch/
- DISC-DNA Information and Stock Center http://www.dna.affrc.go.jp

Final Thoughts

- Trust your intuition
- Look at all the possibilities
- Use all the resources you can

How we develop Prosite patterns!




Brigitte Boeckmann / 1994
"'Potential', 'Probable', 'By Similarity' –
They don't know anything about this protein!"